Social Identity As Hierarchies Of Selves*

Theory and Experimental Evidence From Regional Contributions in Mexico

Mauricio Fernández Duque

March 30, 2017

Abstract

A recent literature in economics has recognized the importance of norms associated to social categories such as nation or region, and of the behavioral impact of priming these categories. However, unexplained phenomena and tensions remain. I develop a model that nests the two main approaches to how primes affect behavior (based on Benjamin et al., 2010 and Bénabou and Tirole, 2011) and the *hierarchies of selves* approach, which assumes that individuals want to signal which norms drive their decisions whether or not they are primed. While a prime may move behavior towards a cognitively accessible norm, it will move it away from the norm if it confounds the signal that another norm motivates behavior. This gives a novel account of how the impact of a prime is affected by situation-specific norms that are not primed. I asked subjects to rank their region and nation. Two weeks later, I randomly assigned them to a national prime and to situations with a national norm. I find support for the unique predictions of the hierarchies of selves approach: the impact of a prime depends on the un-primed regional norm. The results have implications for priming experiments designed to infer norms, problems of replicability and recommendations of when to use norms to affect behavior.

Keywords: identity, norm, priming, pro-social behavior, region, nation

JEL Classification: C91, C92, D03, Z13

*Corresponding author: Mauricio Fernández Duque, Center for Public Leadership, Harvard Kennedy School, Cambridge, MA 02138, USA. Email: duque@fas.harvard.edu. I would like to thank participants at conferences at Harvard, MIT and ITAM. Discussions with and comments with Sam Asher, Nava Ashraf, Iris Bohnet, Ryan Enos, Ben Golub, Michael Hiscox, Torben Iversen, Horacio Larreguy, Chris Lucas and John Marshall were particularly useful. Generous support was provided by the Harvard Experimental Working Group. José Ignacio Ávalos and his team at Un Kilo de Ayuda, as well as Diego Domínguez and Vidal Romero, were very helpful in the implementation of the experiment. Alexandra Uribe made edits and suggestions that enriched the paper. All errors are mine.

1 Introduction

To affect behavior, policymakers, marketers and researchers send persuasive messages. These messages are often uninformative – they do not transmit any private information that would change a rational actor's optimal choice. However, uninformative messages can change behavior through *priming*, or making some thoughts more cognitively accessible. An important instance of this type of persuasion is priming social categories such as race, religion or nation. There is much experimental work that shows that priming social categories can affect behavior (Cadsby et al., 2013; Boschini et al., 2014; Transue, 2007; Chang et al., 2015; Chen et al., 2014; Cohn et al., 2014, 2015; Pedic, 1990; Reed, 2004; Bertrand et al., 2010; Berger et al., 2008) although it has failed in some contexts and there have been problems with replication of results (Fryer et al., 2008; Open Science Collaboration, 2015; but see Gilbert et al., 2016). Several approaches have been advanced to explain the impact of primes (Bénabou and Tirole, 2011; Benjamin et al., 2010), but the literature has lacked a systematic approach to empirically distinguish between them (Klein, 2014; Cohn and Maréchal, 2016).[1] In this paper, I develop a simple model that nests three approaches of how primes affect behavior, derive a test to distinguish between the approaches, and provide experimental evidence in favor of a novel approach which I call *hierarchies of selves*.

To fix ideas, consider a situation in which an individual is deciding whether to make a charitable donation to a recipient of the same nation. We have much experimental evidence that there is a norm to benefit those of a shared group (Akerlof and Kranton, 2005, 2010; Chen and Li, 2009). This norm may become more cognitively accessible by priming the nation, for example with subtle displays of national symbols.[2] According to the 'framing' approach, individuals react by increasing

[1] The work on this topic in social psychology is vast and hard to summarize. Social identity theory is an umbrella term for a set of theories that broadly state that the social categories an individual belongs and the emotional significance they attach to them affects their behavior (e.g. Turner and Reynolds, 2012; Tajfel, 1981; Tajfel and Turner, 1979). Swann and Bosson, 2010 respectively refer to the two main sub-theories of the social identity theory as the 'self-enhancement' and 'self-categorization' approaches. Hoff and Pandey, 2014 make a similar distinction by dividing social identity theories into fixed self and frame dependent self, as do Guala and Filippin, 2016. These sub-theories provide different predictions of the impact of primes on behavior.

[2] The view that primes impact behavior by making norms more cognitively accessible is argued by Akerlof and Kranton (2010); Benjamin et al. (2010); Bénabou and Tirole (2011); Guala and Filippin (2016); Hoff and Pandey (2014). Although norms are not always easy to define, they are well understood in some situations such as the ones

donations since they automatically think their behavior should follow the national norm.[3] An alternative 'signaling' approach is that the prime moves individuals to signal their attachment to the nation.[4] Those who are strongly attached to the nation react to the prime by increasing charitable donations, but those without a strong attachment may not react to the prime or even react negatively by decreasing the donation.

Now consider two versions of donating to a co-national: the recipient and decision-maker also share the same sub-national region, or they are of different regions. From the point of view of the framing and signaling approaches, the impact of priming the nation (without priming the region) is identical in either version, since what matters is simply that there is a national norm to donate to a co-national. However, in the first situation there is a regional norm to donate absent from the second, since the recipient and decision-maker are of the same region. Whether there is a regional norm when the nation is primed will matter in the hierarchies of selves approach.

I develop a model that encompasses the three approaches. An individual belongs to two social categories, a region and a nation, and has private information about how strongly she values each. In a given situation, a behavior may be expected of those who value and belong to a social category.[5] For example, donating to someone from a specific region is expected of those who value and belong to the region, but less so of others. In that situation, we say that donating is a norm associated to

we will consider.

[3]The framing approach corresponds to the 'self-categorization' or 'frame-dependent self' theories of footnote 1. I call it 'framing' because the the prime 'frames' the situation in terms of the norm. Experiments that provide evidence in favor of the framing approach include Shih et al., 1999; Benjamin et al., 2010, 2016; Gaertner and Dovidio, 2012; LeBoeuf et al., 2010; Charnysh et al., 2014; Transue, 2007; Sachs, 2009; Afridi et al., 2015; Berger et al., 2008; Cadsby et al., 2013; Chen et al., 2014; Cohn et al., 2014, 2015; Dee, 2014; Chang et al., 2015; Cohn et al., 2015. These experiments show that primes move behavior in the direction of a norm on average.

[4]The self-enhancement sub-theory mentioned in footnote 1 broadly claims that individuals make identity-related choices to maximize their self-esteem. A fruitful interpretation of this self-esteem maximization is a process of signaling – to others or to oneself (Bénabou and Tirole, 2011). Experiments that favor the signaling approach include Reed, 2004; Morris et al., 2008; Pedic, 1990; Seiter and Gass, 2005; Ledgerwood and Chaiken, 2007; Martin et al., 1984; McKay et al., 2011 . These experiments find that the impact of a prime is stronger among those who are more attached to the primed social category, or among those who act in other ways consistent with its norms.

[5]Having a behavior expected of someone implies an audience. The paper is agnostic as to whether this audience is real or imagined. In an anonymous experimental setting, this audience could be the individual herself, or the experimenter.

the region, or a 'regional norm'. A social category may or may not have an associated norm in a given situation. There is no regional norm to donate to someone from a different region, although there is a national norm to do so. When there is no regional norm, there is no behavior expected of those who value and belong to the region. Table 1 summarizes the regional and national norm for the two situations we've discussed.

How individuals value their social categories and the norms in a given situation are primitives in the general model. A third primitive is whether the situation includes a national prime. We will want the model to make predictions of whether a national prime moves behavior 'towards' or 'away' from a norm, and about 'where' behavior 'goes'. Therefore, the model assumes there is a distance between behaviors. The absolute difference between donations is a natural candidate in the situations we're considering. The model further assumes a default behavior, which is the ideal behavior for those not motivated to follow a norm. In our example, the default behavior is a small donation. We can then ask in which situations and for which types the national prime will move behavior away from a large donation (the national norm) towards a small donation (the default behavior), and in which situations and for which types it will move behavior away from a small donation and toward a large donation.

Individuals get utility from following the cognitively accessible norm, which is determined by the three primitives: without a prime it is determined by how they value social categories, but with a prime it is determined by the prime. In the framing approach, individuals follow the default behavior unless there is a cognitively accessible norm. In the signaling and hierarchies of selves approaches, individuals get additional utility from what their behavior signals when the nation is primed.[6] In the signaling approach, individuals signal how strongly they value the nation. In the hierarchies of selves approach, they signal how they value the nation relative to the region – individuals avoid a norm if it sends the unwanted signal that they value the primed nation more than the un-primed region.

Table 2 summarizes the impact of a national prime on regionalists – those with a high value of the region and a low value of the nation. According to the hierarchies of selves approach, the impact of the national prime depends on both the national and regional norms of a situation.

First consider the decision to donate to someone from the same region. In this situation, there

[6]Signaling implies an audience. The audience is as discussed in footnote 5.

is both a regional and national norm to give a large donation. Without a prime, the regional norm is cognitively accessible to a regionalist, so she gives a large donation. A national prime makes the national norm cognitively accessible. Since the national and regional norm are the same, giving a large donation when the national norm is cognitively accessible sends an unwanted signal about what a regionalist values. Regionalists react by avoiding a large donation. The default behavior (a small donation) is the preferred way to send this signal – as opposed to, say, a costly alternative such as giving an even larger donation. Therefore, the national prime decreases donations in this situation.

Now consider the decision to donate to someone from a different region in the same nation. There is a national norm (a large donation) but there is no regional norm. Without a prime, the national norm is not cognitively accessible to a regionalist, so she follows the default behavior (again, a small donation). Since there is no regional norm, following a national norm does not send an unwanted signal. Therefore, the national prime moves behavior towards the national norm. In this situation, the national prime increases donations.

The theoretical results show that a prime may therefore move behavior in the direction of the primed norm, but in some situations it will backfire. The backfiring effect of a prime has been observed in (Pedic, 1990; Gómez et al., 2008; Bhattacharjee et al., 2014; Thomas et al., 2015; Puntoni et al., 2011).[7] The model makes a sharp prediction of when there will be backfiring: a national prime backfires for regionalists in a situation with the same national and regional norm.[8] These results can also explain problems of replicability. Unobserved variation in how subjects value social categories means that for some situations, a prime will sometimes move average behavior

[7] A summary of the explanations given in these papers of why backfiring occurred in these papers was that individuals were signaling that the prime did not reflect how they valued their social categories. For example, Thomas et al. (2015) argues that primes backfire when they showcase an unfavorable view of an individual's social categories. This explanation is similar to the hierarchies of selves approach.

[8] Notice that not all situations with both a regional and a national norm would lead to backfiring. In Section II, I show that in situations when the national and regional norms are different, the prime does not lead to backfiring and may actually move behavior towards the primed norm. As an example of a situation where the national and regional norms are different, I consider a situation in which a subject is ranking her attachment to her nation and to her region. I will argue it is an example with both regional and national norms where a national prime would move regionalists' behavior towards the national norm. I provide experimental evidence for this prediction in the empirical section.

towards the norm, and sometimes away from it.

Past approaches have ignored the impact of norms that are not primed. This can be seen in Table 2 for the framing and signaling approaches, whose predictions do not depend on the situation. However, we should be concerned that when we add new variables to theories – in this case how social categories are valued relative to each other, and which social categories prescribe a norm in a situation – we are overfitting. This is a concern if we're explaining n stylized facts with n new variables. The current approach does explain why primes may backfire, and how unobserved heterogeneity may explain problems of replicability. However, it also makes sharp and novel predictions regarding how regionalists will react to primes across situations where the unprimed norm differs. These novel predictions have not been previously tested. To the best of my knowledge, this is the first work that empirically studies how relative attachment to social categories affects how primes impact behavior. Further, by looking at these heterogenous effects I will be able to test between the signaling, framing and hierarchies of selves approaches. Table 2 shows that donating to a co-national of the same region and to a co-national of a different region are sufficient situations to distinguish between the three approaches.

Mexican students participated in a two-round online experiment. In the first round, students were asked to select the states in Mexico that formed their sub-national region, and were asked to simultaneously rank their attachment to their region and to the nation. We used the answers to this baseline, which did not contain a prime, to classify 'regionalists' – those who ranked the region higher than the nation. Two weeks later, they filled out an endline survey. They were exposed to a donation petition benefiting a state either in their region or outside their region. A national prime was included in some donation petitions. In addition, we asked them once again to rank their attachment to their region and to the nation.

The results are consistent only with the hierarchies of selves approach. A prime decreases regionalists' donations to those of a shared region, but increases donations to those of a different region. Consistent with past failures at replicating priming experiments, results that do not disaggregate by regionalism are insignificant. Results for those not classified as regionalists are also consistent with the theory.

The survey provides further evidence supporting the causal mechanisms behind the hierarchies of selves approach. To ensure that the prime was uninformative, we asked a battery of questions

about the NGO making the donation petition, as well as the state receiving the donation. The prime was not predictive of these responses. Subjects were, however, more likely to mention Mexico in a free-word association task, suggesting that the prime made the nation more cognitively accessible. Finally, we provide some suggestive evidence that regionalism affects cognitive accesibility: regionalists are more likely to think of the beneficiaries as members of a region without a prime, but this difference disappears with a prime.

Recent papers have attempted to infer the norm associated to a social category through priming experiments (Benjamin et al., 2010, 2016; Cohn et al., 2014, 2015; Cohn and Maréchal, 2016). The difference in behavior with and without a prime is used to infer the norm. The framing approach has been used to justify these norm-inferring experiments, since primes always move behavior towards the norm. However, competing approaches challenge what we can infer from these designs, and point to the designs that would be needed to properly infer these norms. In order to advance on the task of norm inference, we need to sort between these competing approaches. Other papers have attempted to distinguish between framing and other approaches by testing whether primes impact behavior in different circumstances (Hoff and Pandey, 2014; Guala and Filippin, 2016). However, they do not compare between alternative approaches of when primes affect behavior. By systematically comparing approaches, this paper contributes to the positive question of what we can learn from primes' impact on behavior, and to the normative question of when to use primes to affect behavior.

The rest of the paper is organized as follows: in Section II, I present the model. Section III presents the experimental deign. Section IV presents the results of the experiment, and Section V concludes.

2 Theoretical Framework

2.1 Setup

This section presents the theoretical model. The model is kept as simple as possible subject to allowing us to test between different approaches to how primes affect behavior.

Consider an individual that can be classified along two social categories. Although the model is applicable to any social categories, I will focus on the case studied in the experiment: individuals

are part of a nation N, and of a region R within the nation, with $c \in \{N, R\} \equiv C$. Individuals are of two types: regionalists ($\theta = R$) and nationalists or non-regionalists ($\theta = N$), with $\Theta \equiv \{R, N\}$ and $P(\theta = N) = \mu$. These types are an individual's 'identity', which is defined by how strongly attached an individual is to her social categories. In order for these types to be interpretable in terms of the signaling or hierarchies of selves approach, we assume that regionalists value the region highly and place low value on the nation, while nationalists value the nation highly and place low value on the region.

After observing his type, the individual makes a decision $d \in D \subset \mathbb{R}$ in situation $s \in S$. There is a norm function $n : S \times C \to D$ that sets the behavior in situation s that is expected of individuals that belong to social category c. Let $d = 0$ be the default behavior, or the preferred behavior of an individual unconcerned with following norms, and interpret $n(s, c) = 0$ as c not setting a norming in situation s. Situation s is accompanied by a message m that may contain a national prime ($m = N$), or not ($m = \theta$). I will only consider situations s where the primed social category sets a norm: $n(s, N) \neq 0$.

After the individual makes his decision, an audience observes d, s and m and updates the individual's type. In an anonymous experimental setting, the audience could be the experimenter or the individual self signaling (Bénabou and Tirole, 2011). Utility is given by:

$$u(d; s, \theta, m, h_s) \equiv -(n(s, m) - d)^2 + \mathbb{1}_{\theta \neq m} h_s P(\theta \neq m | d)$$

where $h_s \geq 0$ is a situation-dependent weight on the audience's posterior beliefs. The interpretation is that the prime makes subjects automatically process the situation in terms of the cognitively accessible national social category. This leads some to signal their attachment and others to follow the national norm. Who does which will characterize the framing, signaling and hierarchies of selves approaches.

Definition 1. *The framing approach is captured by $h_s = 0$. The signaling approach is captured by $h_s \propto |n(s, m)|$. The hierarchies of selves approach is captured by $h_s \propto |n(s, \theta)|$.*

When $h_s = 0$ for all s, signaling considerations are irrelevant. Individuals follow the norm of whichever social category is cognitively accessible, be it the one they are most strongly attached to or one that is primed.

When $h_s > 0$, individuals become interested in signaling their identity when the primed social category is different from their identity ($\theta \neq m$). If further $h_s \propto |n(s,m)|$, then these signaling concerns arise whenever the primed social category norms a situation. This is the signaling approach.

The hierarchies of selves approach assumes that signaling concerns depend on social categories that are *not* primed ($h_s \propto |n(s,\theta)|$). Those for whom the primed social category is low on their hierarchies of selves want to reveal their type only when the situation is normed by a social category that is higher on the hierarchy. Otherwise, individuals follow the norm of the primed social category. Below I derive conditions under which these approaches can be distinguished.

I am interested in Perfect Bayesian Equilibria. I would like to use the D1 criterion to refine out-of-equilibrium beliefs. To do so, I turn the audience's beliefs into their strategy. For example, they get utility of 1 from guessing the probability that $\theta \neq m$ given the individual's decision, 0 otherwise, and they are risk neutral.

2.2 Examples

I now capture the three examples discussed in Section I, and which will be used in the experimental setting. First, in $s = in$ an individual makes a donation decision to a beneficiary in their region. I assume that $D = \{0, 1\}$ and $n(in, R) = n(in, N) = 1$. That is, donation could be 'small' ($d = 0$) or 'large' ($d = 1$), and both the regional and national norm are to give a large donation. Second, in $s = out$ an individual makes a donation decision to a beneficiary outside of their region but within their nation. I again assume $D = \{0, 1\}$, but now $n(in, R) = 0$ and $n(in, N) = 1$. Although there is still a national norm to give a large donation, there is no longer a regional norm. As previously discussed, we have much evidence on individuals' in-group bias, which is frequently interpreted as arising from a norm to act favorably to in-group members (Akerlof and Kranton, 2005, 2010; Chen and Li, 2009).

Third, in $s = att$ individuals are asked to rank their attachment to the region and to the nation. This situation differs in that expected behavior depends on how one social category is valued relative to another. The norm for regionalists in this situation is to rank the region more highly, and the norm for nationalists is to rank the nation more highly. This can be captured by assuming that $D = \{-1, 0, 1\}$, $n(att, R) = -1$ and $n(att, N) = 1$. The interpretation is that $d \in D$

is the difference between strength of attachment to the nation versus the region. The intermediate behavior is to rank the region as high as the nation. This is $d = 0$ in our example. It is hard to justify $d = 0$ as a 'default behavior' in the sense we've done for our other examples. In order for our arguments to go through, this interpretation will not be necessary in this situation. All that will be needed for our results is for there to be an intermediate behavior between the national norm and the regional norm. We have set the behavior to be $d = 0$ for convenience.

In the introduction, we defined a norm as a behavior that is expected of those who value and belong to a social category. However, the donation and ranking situations differ in what the verb 'value' in the definition refers to. In the donation situations, donating a large amount is expected of those who have a high absolute value for those who share a social category with the recipient. In the ranking situations, ranking one social category higher is expected of those who value one social category highly relative to another. This distinction mirrors the difference between the signaling and hierarchies of selves approaches: types in the signaling approach are defined by whether they have a high and low absolute value of a social category, while types in the hierarchies of selves approach are defined by how value of a social category relative to another. In the model, the distinction does not matter since we've defined types as having both high absolute and relative value for a social category. Therefore, the norms in all the situations we've discussed apply to either approach. This is restrictive for the hierarchies of selves approach if there are subjects who who value both social categories the exact same amount. We assume that there is a negligible amount of individuals who do so.

2.3 Relationship to Past Models

In this sub-section and the next, I discuss the relationship of the model to the ones it captures as special cases. They can be skipped without loss in continuity.

The seminal work on integrating social identity into economics is Akerlof and Kranton, 2000. They argue that own and others' actions are normed[9] in specific situations by their social categories, which in turn affect behavior. They also argue that individuals' given characteristics may affect

[9]In Akerlof and Kranton, 2000, they use the term 'prescriptions' instead of 'norms', arguing that 'norms' has taken on many alternative meanings in the economics literature. They revert to the use of the term 'norms' in Akerlof and Kranton, 2010, a practice which I follow.

the impact of following the norms of a given identity.[10] This line of research does not explicitly take into account the impact of primes on the cognitive accessibility of these norms. Therefore, a simplified version of their model can be captured by $u_{AK} = -(d - n(s,\theta))^2$.[11] Trivially, one could adjust this model to take into account primes by defining a situation with and without a prime as different situations. However, this makes for a poor theory: it is hard to derive testable predictions across situations when each tweak on a situation defines a new set of norms. For this reason, primes are typically thought of as independent from situations.

The framing approach ($h_s = 0$) is captured by Benjamin et al., 2010, 2016. A simplification of their utility function can be obtained through a slight modification of the model:

$$u_{BCS}(d; s, \theta, m, d_0) = -\mathbb{1}_{m\neq\theta}(d_0 - d)^2 - \mathbb{1}_{m=\theta}(n(s,N) - d)^2$$

They assume $h_s = 0$, remain agnostic about the value of d_0, and use a prime to estimate the norm associated to the primed social category in a given situation, $n(s,N)$. This norm is assumed to be equal for all individuals that share a social category. In contrast, I take a stand on what the norm of each social category is in a specific situation, allowing it to differ by type ($n(s,\theta)$) and use this to estimate h_s. The force of the argument in this paper thus strongly depends on correctly assessing what behavior is normed by a social category in a given situation, which is why I have chosen situations where norms are well understood.

The signaling approach ($h_s \propto |n(s,m)|$) is closely related to the self-signaling model of Bénabou and Tirole, 2011. A simple version of their utility function can be written as follows:

$$u_{BT}(d; s, \theta, m, h_s) = \mathbb{1}_{m=N} h_s P(\theta = N|d) - (n(s,\theta) - d)^2$$

Their model differs in that all individuals are assumed to want to signal the same identity – what they refer to as the *good identity convention*. It further differs in that they assume nationalists are more responsive to the national norm with and without a prime, which serves to fulfill the

[10] Follow-up work has used these characteristics to endogenize strength of attachment to social categories (Wichardt, 2008; Shayo, 2009).

[11] The utility function in their paper takes into account the impact of others' behavior on identity, but I focus on decisions in the experiment that are independent of others' actions.

single-crossing property. For this to hold with u_{BT}, it must be the case that $n(s,R) \neq n(s,N)$. Despite the differences, consider the following:

Proposition 1. *Suppose* $2n(s,R) - \min D < n(s,N) = \max D$. *Then when* $m = N$, *only two pure strategy equilibrium can be sustained for* u_{BT} *and* u *with* $h_s \propto |n(s,m)|$. *There is a pooling equilibrium where both types choose* $n(s,N)$, *and a separating equilibrium where* N *types choose* $n(s,N)$ *and* R *types' choice is closer to zero than* $n(s,N)$.

Proof. In the appendix. □

What drives individuals to separate according to u_{BT} is that the type who highly values the primed nation is more motivated to act according to its norms. In the signaling approach in the model, separation is driven by those with a low value of the primed nation who want to reveal their type. Proposition 1 suggests that either specification makes very similar predictions in a range of parameter values.[12] The restriction on the parameter values can be relaxed, but allows me to state the result simply without the need for additional notation.

2.4 Definitions of Identity

It is worth noting what the term 'identity' has referred to in the theories we've mentioned. For Akerlof and Kranton, 2000, identity is the utility derived from how actions and characteristics relate to the norms of an individual's social categories. However, they claim that identity also refers to social categories. Benjamin et al., 2010, 2016 do not pinpoint what they refer to as identity in their model, but state that it is inspired by Akerlof and Kranton, 2000. In contrast, for Bénabou and Tirole, 2011 identity is an individual's type, which captures the strength of attachment to an identity. However, they do not consider social categories in their analysis. I conclude this discussion by noting, first, that the term 'identity' has been used to refer to different parts of complementary formalizations. Second, I have maintained a distinction between social categories, how these social categories are valued (what I refer to as identity), and the utility derived from following norms associated to a social category. I believe this terminological distinction adds clarity, but is ultimately a hybrid of past conceptualizations.

[12] As we show in the proof of Proposition 1, the only difference between the signaling approach in the model and u_{BT} comes from the behavior of R with a prime. According to u_{BT}, R types choose $n(s,R)$. According to the signaling approach, they choose the decision closest to $n(s,N)$.

2.5 Results

First, I will establish a straightforward result regarding behavior when signaling concerns are absent. Let $d^*(s, \theta, m, h_s) \equiv \arg\max_{d \in D} u(d; s, \theta, m, h_s)$.

Lemma 1. *$d^*(s, \theta, \theta, h_s) = n(s, \theta)$ and $d^*(s, N, N, h_s) = n(s, N)$.*

Trivially, individuals follow the norm of their identity without a prime, and nationalists follow the national norm with a prime. This implies that I need only focus on the decision of type R with a prime to distinguish between the theoretical approaches. Despite the straightforward result, note that it makes different predictions for the signaling and hierarchies of selves approaches given how types are defined. As previously mentioned, types are defined in terms of attachment to a single social category in the signaling approach, and as relative strength of attachment in the hierarchies of selves approach.

The next result narrows down the types of decisions the R type makes with a prime. Let $d^c \equiv \arg\max_{d \in D \setminus n(s,N)} -(n(s, N) - d)^2$ be the set of decisions that are closest to the national norm.

Lemma 2. *If $h_s \geq (d^c - n(s, N))^2/\mu \equiv h^*$, $d^*(s, R, N, h_s) \in d^c$. If $h_s < h^*$, $d^*(s, R, N, h_s) = n(s, N)$.*

Proof. This is shown in the proof of Proposition 1.□

Lemma 2 tells us that a national prime will either push behavior towards the national norm, or to behavior that is close but distinct from the national norm.

With these preliminaries in hand, we can move to the first set of results. Let

$$G(s, h_s) \equiv (d^*(s, R, N, h_s) - n(s, N))^2 - (d^*(s, R, \theta, h_s) - n(s, N))^2$$

$G(s, h_s)$ is the change in the gap between what the R type does and the national norm with and without a prime. I am interested in characterizing when the prime affects the gap.

Result 1.

- *Suppose $n(s, R) = n(s, N)$. Then if $h_s \geq h^*$, $G(s, h_s) > 0$. If $h_s < h^*$, $G(s, h_s) = 0$ with $d^*(s, R, m, h_s) = n(s, N)$.*

- *Suppose* $\exists d \in (n(s,R), n(s,N))$ *(assuming, without loss of generality, that* $n(s,R) < n(s,N)$*). Then* $G(s,h_s) < 0$*.* $G(s,h_s) < G(s,h_s')$ *if* $h_s \geq h^* > h_s'$*.*

- *Suppose* $\nexists d \in (n(s,R), n(s,N))$*. Then if* $h_s \geq h^*$*,* $G(s,h_s) = 0$*. If* $h_s < h^*$*,* $G(s,h_s) < 0$*.*

Proof. This can easily be verified by applying Lemmas 1 and 2.□

Result 1 shows that if $h_s \geq h^*$, we must take into account the multiple social categories that norm a situation to understand behavior. A prime may move indviduals towards or away from the norm of the primed social category. The gap closes only if the regional and national norm are different: $n(s,R) \neq n(s,N)$. This poses a challenge to experimental designs that use primes to infer the norm of a group (e.g. Benjamin et al., 2010, 2016; Cohn et al., 2014, 2015), since this inference will lead to wrong conclusions if $n(s,R) = n(s,N)$ and $h_s \geq h^*$. Having to know whether multiple social categories norm a situation further puts these designs in a tight spot since their objective is to estimate a single norm.

One approach to estimating a norm that is suggested by the framework is to compare behavior among those who value a social category highly (relatively and absolutey) with those who place a low value on a social category. This approach presents its own challenge of looking for an exogenous shock to how a social category is valued. Alternatively, we can observe the behavior of those who highly value a social category with and without a prime.[13] As Benjamin et al., 2016 recognize, this approach has the limitation that those with a strong attachment may already be following the norms of their identity.[14] This is not the approach I follow in this paper. I will look at settings where norms are arguably well understood, and exploit Result 1 to test for different theories of h_s.

Corollary 1. *Let* $S = \{in, out\}$*.*

- $h_s < h^* \ \forall s \in S \Leftrightarrow d^*(s,R,N,h_s) = n(s,N) \ \forall s \in S$

- *If* $d^*(s,R,N,h_s) \neq n(s,N) \Leftrightarrow n(s,N) \neq 0 \ \forall s \in S$*, then* $h_s \propto |n(s,m)|$*.*

[13] Although the simple model predicts that they will behave the same way with and without a prime, a natural modification to the model would allow individuals to be biased towards a default action different from the normed behavior.

[14] The proposed approach differs from past work that suggests norms be identified through a coordination game, with the logic that the norm would serve as a focal point (Krupka and Weber, 2013). Since their procedure selects a norm for any situation, it is not compatible with the key tenet of the design that some situations are not normed by a social category.

- *If* $d^*(s, R, N, h_s) \neq n(s, N) \Leftrightarrow n(s, R) \neq 0 \ \forall s \in S$, *then* $h_s \propto |n(s, \theta)|$.

Corollary 1 shows us how we can separate framing, signaling and hierarchies of selves by looking at how individuals behave with the national prime across two situations: the decision to favor someone of the same nation who respectively is part (*in*) or is not part (*out*) the same region. Past studies have looked at similar situations, replacing a common region with a common race (Transue, 2007) or religion (Charnysh et al., 2014). Two stylized facts from these experiments is a baseline bias towards an in-group without a prime and a reduction of this bias with a prime driven by an increase in donations to the out-group. Lemma 1 and Corollary 1 show that these stylized facts favor the framing approach over the signaling approach. Regionalists would decrease their donations according to the signaling approach, diminishing average contributions. The results further show that *in* and *out* are sufficient to test for the hierarchies of selves approach: it makes the unique prediction that regionalists' donations will decrease with the prime in the in-region, but increase in the out-region.[15] However, since past studies did not look separately at subjects with different identities, this prediction has not been tested.

Situation $s = att$ helps us further test a prediction of the hierarchies of selves approach. Consider the following:

Result 2. $h_s \geq \mathbb{1}\{|n(s, \theta)| > 0\}h^* \Leftrightarrow$ *in equilibrium,*

- $G(s, h_s) > 0 \Leftrightarrow n(s, N) = n(s, R)$
- $G(s, h_s) < 0 \Leftrightarrow 0 = n(s, R)$, *or* $\exists\ d \in (n(s, R), n(s, N))$ *with* $0 \neq n(s, R) \neq n(s, N)$
- $G(s, h_s) = 0 \Leftrightarrow \nexists d \in (n(s, R), n(s, N))$ *with* $0 \neq n(s, R) \neq n(s, N)$

Proof. In the Appendix. □

[15] If I modified the choice set from $D = \{0, 1\}$ to contain more choices between $n(out, R) = 0$ and $n(out, N) = 1$, the signaling and hierarchies of selves approaches would look more similar in situation $s = out$ by Result 1. Indeed, according to the signaling approach in the model, the R type would choose the donation that is closest to $n(s, N)$, while in the hierarchies of selves approach they would choose $n(s, N)$. However, this is a result of having made the atypical assumption that those who place a low value on the nation are motivated to signal their type. In Bénabou and Tirole (2011), captured by u_{BT} in section 2.3, it is those who place a high value on the nation who are motivated to signal their type. As shown in the proof of Proposition 1, in a separating equilibrium with utility u_{BT}, regionalists would choose $n(s, R)$.

Result 2 characterizes when a prime will increase normed behavior with a hierarchies of selves approach. A national prime has two conflicting effects on regionalists. On the one hand, they are motivated to signal their identity by choosing behavior that is distinct from the national norm. This increases the gap between their behavior and the national norm if the national and regional norms are the same, such as in $s = in$. On the other hand, they are primed to behave according to the national norm. If there is an action that moves them closer to but is still distinct from the national norm, the prime will move them towards that action. This decreases the gap, such as in $s = att$.

This Result shows that in some situations that are normed by an individual's identity ($n(s, \theta) \neq 0$), priming a different social category may move behavior closer to its norms. This result may help explain why primes have had mixed success with replication, and helps us guide policy decisions of when to use primes to impact behavior. By using Lemmas 1 and 2, it can be verified that the policy recommendations would differ for the signaling and framing approaches. This can be seen with the three situations I consider.

Corollary 2. *$h_s \propto |n(s, \theta)|$ and $h_s \geq h^* \Leftrightarrow$ in equilibrium, $G(in, h_{in}) > 0$, $G(out, h_{out}) < 0$, $G(att, h_{att}) < 0$*

The result from Corollary 2 is summarized graphically in Figure 1.

3 Experimental Design

The experimental design is summarized in Table 3. It consisted of a baseline survey where subjects were first asked which state they were most attached to. They then selected the states that were part of the same region, and indicated their attachment to that region and to the nation. In the endline, subjects encountered a donation petition asking to help beneficiaries of the NGO. I block randomized on relative attachment - the difference between regional and national attachment - to assign subjects beneficiaries in or outside their region, and cross randomized whether the petition contained a prime.

After being exposed to the donation petition, they were asked whether they would like to donate part of their participation fee, a battery of questions about the NGO and the state, and were again asked about their regional and national attachment. The attachment questions were asked

simultaneously, so the subjects are effectively asked about their relative strength of attachment. The in-region and out-region donation decisions, along with the questions about attachment, constitute the three situations I use to distinguish between different approaches to how priming social categories affect behavior. Other questions were included to test for cognitive mechanisms.

In November 2015, the approximately 5,000 students at ITAM were sent four email invitations in the space of two weeks to take part in a two-part study consisting of a baseline survey and an endline survey. They were not offered a fee for taking part in the baseline experiment, and did not know they would be offered a fee to take part in the endline. 955 subjects participated in the baseline survey, and 580 of those returned for the endline. Both the baseline and endline were designed and run through Qualtrics, and each took approximately ten minutes to complete.

The baseline survey had five categories of questions: Demographic characteristics, political views, views on charitable giving, perception of state income, and questions to determine regionalism. This last category was determined in the following manner: I first asked them to indicate which state they felt most strongly attached to. I then determined their region by asking them to indicate on a map which states they considered to be part of the same region as the state they chose. I avoided using an exogenous division of Mexico into regions since 1) there is no clear consensus on how to do so (Liverman and Cravey, 1992), and 2) I obtain the most statistical power by defining individuals' in-region and out-region the way they would define it. With few exceptions explained below, I use region as the students defined it. After asking about their region, I ask them about their attachment to their nation.

To randomize, at the beginning of the baseline survey I assigned individuals a state in their region or outside their region. I further randomized whether the state was rich or poor. The rich (poor) states the individuals were assigned were subjectively identified by the individual as rich (poor), and either widely perceived by others to be rich or objectively rich (poor) using 2013 GDP. I use this to control for donations based on need, and not identity.

13% of respondents' in-region states did not vary in terms of income (i.e. having both rich and poor states). I faced a trade-off between sticking strictly to the states the individual selected as part of his or her region and balancing on state income. I favored the latter. If a region had a middle-income and rich or poor state, I considered it balanced, if not, I added a state that at least a third of the individuals of the same state considered part of their region. In practice, only 5%

of those assigned a state in their region were assigned states in their expanded region. Once I had defined the set of in-region rich and poor states and out-region rich and poor states I randomly chose one state from each set. I then assigned individuals to one of four conditions: in-region with prime, in-region without prime, out-region with prime, and out-region without prime.

I block-randomized on regionalism. Regionalism is defined by the difference between their reported attachment to their region and their nation.

Between two and three weeks after they filled out the baseline survey, I invited subjects to fill out the endline survey. I announced that I would raffle several 100 peso prizes (approximately 7 US dollars at the time) with about a 10% chance of winning. Subjects were told that they would be exposed to the design of a card used for fund-raising, some of which were used in the field. Examples of these cards are shown in Figure 2.

The cards feature the name of the NGO, three children and the caption 'Thanks for nourishing the children of [name of state]'. The card varies on two dimensions: the name of the state, and whether it contains the colors of the Mexican flag. There are 32 states in Mexico, so there were 64 cards total. The nationalist colors frame the cards, and appear on the name of the state. The control condition uses teal color instead. Subjects were exposed to the design for only ten seconds, to avoid excessive analysis of the card.

The endline survey had three types of questions.

1. *Benefits to recipients in their assigned state.* Subjects played a dictator game in which they could allocate any part of the proceeds from the lottery to beneficiaries in the recipient state. A question on redistribution was also asked, but it was found to be too vague and I will not delve into the answers.

2. *Regional and national attachment*: These are the same questions as in the baseline.

3. *Cognitive mechanisms*: I included a set of questions to test three claims about the mechanism behind the prime's impact: that the prime was uninformative, that the prime made the nation more cognitively accessible, and that regionalism affects cognitive accessibility. A prime works because it makes a pre-existing concept or idea more cognitively accessible. Therefore, a prime is uninformative as it does not give more information, it just reinforces an idea or concept already held. Otherwise, the mechanism behind a prime's impact is confounded.

The first two categories are the main dependent variables. The third provides evidence of causal mechanisms.

4 Summary Statistics

Table 4 shows summary statistics. About 59% of subjects are male, 82% are between 17 and 21, and 37% are white as opposed to indigenous or mestizo. 45% of subjects trust the NGO 'a lot' or 'some' and 51% make contributions. Importantly, the state that almost half of the sample is most attached to is Mexico City. The second closest state is the State of Mexico, which shares a border with Mexico City and 9% of the subjects most attached to it. The third most attached to state is Nuevo Leoón, a relatively rich state from the north, with 4% of students. The rest of the sample is dispersed across all other states.

Table 5 displays baseline regional and national attachment. Notice that 481 out of 587 subjects responded 4 or 5 on the five-point scale regarding how attached they are to the region or the nation. This suggests that the attachment to the region or the nation is relatively high among subjects, and very few people have a difference of more than 2 points between attachment to region and attachment to nation (3.7%). Regionalists - those who are more attached to their region than to the nation on a Likert scale, are 15% of subjects (87 subjects total). Nationalists, defined analogously, are 50%. Although nationalism drops by 11 percentage points in the baseline, it is worth noting that of the 555 subjects who completed the second survey, 51% had been classified as nationalists in the original baseline.

Summary statistics of the dependent and independent variables can be found in Table 6. The variable 'Poorer' is equal to one if the subject was assigned a poor or relatively poor state. I found 61% of individuals found it easier to think of their state as part of a nation, rather than a region, but only 31% considered their state to be a state 'like others'.

On average, individuals gave 63 of their 100 pesos to the NGO beneficiaries. Three modes comprised 82% of observations: 52% of subjects gave their full endowment, 16% half, and 14% zero. In terms of the model in Section II, we can think of giving the full endowment as a 'large' donation, and the rest as 'small' donations.

Balance tables can be found in the appendix. Overall, treatments were balanced, as was at-

tachment.

5 Empirical Specification

The empirical specification I will use throughout is the following:

$$Y_j = \alpha + \beta IndVar_j + \gamma_1 Donor_j + \gamma_2 Recipient_j + \delta X_j + \varepsilon_j$$

where $IndVar_j$ is a vector of independent variables, $Donor_j$ is a dummy for the state the donor is most attached to (donor state for short), $Recipient_j$ is a dummy for the recipient state and X_j is a vector of controls. I use robust standard errors throughout.

The ideal experiment would randomize the characteristics of states other than region that impact donations, but I am not able to do so. This departure from the ideal experiment is a concern whenever we study natural groups as opposed to groups created in the lab, but we gain in external validity. The design tries to lessen this concern through fixed effects, by looking at donations to and from many different regions, and by balancing the wealth of assigned in-region and out-region states.

6 Results

6.1 Donation Decisions: $s \in \{in, out\}$

The main result is summarized in Table 7, which captures $s = in$ and $s = out$ from Section 2. The results were estimated from a regression with controls, and a specification without controls yields a similar result. The prime significantly decreases in-region donations for regionalists ($p = .0145$) and significantly increases out-region donations ($p = .02$), while not affecting nationalists donations ($p = .8134$ and $p = .3802$). The difference between the in-region and the out-region impact of the prime is significant for the regionalists ($p = 0.0232$), and insignificant for nationalists ($p = .67$). Comparing the difference in difference of regionalists and that of nationalists yields a significant triple difference ($p = .02$).

Of the eight cells, there are only two conditions in which donations are not statistically higher than 50 pesos at the 10% level: regionalists giving to the out-region without a prime and giving

to the in-region with a prime. I therefore classify the latter two conditions as a small donation ($d = 0$), while the rest were classified as a large donation ($d = 1$).

Table 8 shows the regression results. Both columns include recipient state fixed effects and fixed effects of the state of the donor. The second specification controls for party preference, income level, trust in institutions, contribution to institutions, age, gender, race, low attachment to region and nation, size of region, and baseline and endline location variables and invitation response.

Notice that these regressions drop the part of the sample that reported being equally attached to the region and to the nation. The hierarchies of selves approach assumes that a negligible proportion of subjects values two social categories exactly the same. A question that forced subjects to rank the region or the nation more highly would have helped us to make sharper predictions for these subjects. However, I was interested in comparing results to the signaling approach, where what matters is attachment to a single social category. By asking subjects to simultaneously rank both social categories on a 5-point scale, I was able to test for both approaches. If this way of posing the question was too coarse to classify some of the subjects as regionalists or nationalists, we should get noisier results from this group. In results not shown, we indeed get similar but noisier results by re-classifying nationalists as those who are at least as attached to the nation than to the region.

Table 9 shows the aggregated impact of donating to one's own region, of being exposed to a prime and of the interaction. The only significant effect is a negative in-region bias, which goes away with controls. This overall negative effect may be due to the high concentration of subjects from the relatively rich region that includes Mexico City. Overall, however, I do not find strong or consistent patterns if I don't consider heterogeneity in relative strength of attachment. In results not shown, I do not find significant effects from looking at heterogeneous effects among those with strong versus *absolute* attachment to either the nation versus the region, as would be predicted by the signaling approach. These two results therefore do not support the signaling nor framing approaches, and these negative results are in line with past studies that have not been able to replicate priming experiments (Fryer et al., 2008, Open Science Collaboration, 2015).

6.2 Endline Attachment: $s = att$

The model assumes individuals had to select one of three choices: more attached to their region, equally attached, or more attached to the nation. The regional (national) norm is to select a higher

attachment to the region (nation). In practice, however, subjects chose to rank the region and the nation on a 5-point scale. As discussed previously, this allowed us to test both the signaling and hierarchies of selves approach. Since these questions came at the end of the survey, I further assume that the situation is unaffected by what happened previously. Figure 3 provides evidence consistent with this characterization. The plot was estimated from a regression with controls, and a specification without controls yields a similar result. As we can see, average attachment for region and nation is at the same high level for both regionalists and nationalists no matter their treatment assignment. Variation in responses comes through relative attachment, and is only impacted by the prime. Predictions regarding the hierarchies of selves approach were further confirmed since the prime moves regionalists behavior closer to the national norm, although their behavior continues to be different than of nationalists.

Table 10 shows the results of the right-hand side of Figure 3 in a regression framework. The outcome variable is the difference between the endline attachment to the nation and to the region. The pattern of results and significance are similar across specifications. In both specifications it is at least 5% significant that regionalists are relatively more attached to the region without a prime, and are relatively more attached to the region than nationalists with and without a prime.

6.3 Mechanism Evidence

The motivation for the hierarchies of selves approach to how primes affect behavior can be broken down into the following four claims. First, the prime makes the national social category more cognitively accessible. Second, the cognitive accessibility of social categories differs for regionalists and nationalists without a prime, but not with a prime. Third, the prime is uninformative - it does not lead to a change in beliefs over the optimal decision. Fourth, it instead makes some signal their attachment and others follow the norms of the primed social category. I now provide evidence for the first three claims.

6.3.1 Cognitive Accessibility of the Nation

To test whether the prime made the national social category more cognitively accessible, I asked subjects to say five words that came to their mind after being exposed to the donation petition. I sorted these words into 16 categories. The only significant category to appear was the mention

of Mexico (mentioned by 69 subjects, $p = .005$), and the PRI political party (mentioned by 6 subjects, $p = .097$). The last result may challenge the claim that the prime was uninformative. The prime may have led to updating about the relationship of the state or the NGO with the political party, especially since the colors of the flag are the same as the largest party in Mexico, who was hegemonic for over 70 years in the 20th century. However, the name of the party was mentioned very rarely and I will further show that individuals did not update the state's political characteristics or the NGO's ties with the government.

6.3.2 Cognitive Accessibility of Regionalists Versus Nationalists

After exposure to the donation petition, I asked subjects whether they found it easier to think of states they were assigned as part of a region or as part of a nation. The question was meant to isolate the most cognitively accessible way of thinking about a state.

Figure 4 shows the results. Without a prime, regionalists think of the states more as part of a region than do nationalists ($p = .0394$), but this difference goes away with a prime ($p = .5871$). Surprisingly, nationalists think of states more as part of a region in the presence of a prime ($p = .0126$), although they continue to think of states more as part of the nation on average (54% of the time versus 69% of the time without a prime).

Table 11 shows the same results in a regression framework. The two specifications are as before. The results are similar across specifications.

As a robustness check, I asked subjects how much they thought the state they were assigned was like other states in Mexico. In contrast to the last question, this question relies on information about characteristics of the recipient state and other states in Mexico. The prime should therefore affect the answer to this question only if it is informative. However, it is intuitive that regionalists would find states to be more distinct than nationalists. In results I do not show, I indeed find that regionalists find that states are more distinct, but the prime does not affect subjects' answers.

6.3.3 Uninformativeness

It is important to test for uninformativeness, i.e. that the prime does not change beliefs that lead to different decisions. An informative prime may signal to subjects that the NGO understands something about how to raise funds, which in turn teaches them something about the NGO's

efficiency. Alternatively, it may be informative about the state's characteristics. To rule out these confounds, I asked subjects a series of questions regarding the state, and the NGO. The prime had a significant effect in only two of the thirty-one variables I considered: it significantly predicted that the governor of the state is less likely to be considered an independent as opposed to any other party (p=0.014), and that the state is considered poorer (p=0.077). As for the former, I have no theoretical reason to have expected this, and it may be due to chance. The latter is more concerning. However, it does not correspond with an overall increase in donations – as preference for helping the poorest states would suggest –, the significance isn't high, and I found no systematic difference of perceived poverty between the in-region and the out-region or between regionalists and nationalists.

7 Conclusion

The paper presents a model of how priming social categories impacts behavior, with two standard approaches and a novel approach as special cases. I use this model to formally derive a test between these three approaches, which I execute experimentally in the context of regional and national social categories. The paper provides evidence of the *hierarchies of selves* approach, in which the impact of a prime depends on the norms of social categories that are not primed. Specifically, the paper provides evidence that when the regional and national norm are the same, a prime will move those who value the region more than the nation away from the norm.

Since primes have been used in a wide range of situations, it is important to understand how they affect behavior. The hierarchies of selves approach makes unique policy recommendations for using primes. For example, suppose a policymaker wants to encourage a normed behavior and is choosing whether to prime a social category: she should avoid the prime if multiple social categories norm the same behavior and a large share of the population is more attached to the social categories that would not be primed.

Conversely, there is an increasing number of experiments in economics which compare primed behavior to non-primed behavior to infer norms. The paper contributes to this literature by testing between approaches to how primes affect behavior, which sheds light on what we can learn from these experiments. The hierarchies of selves approach implies that priming a social category will

not always allow us to infer a norm, since it may lead some individuals to move away from it. If we think of the prime as an instrumental variable, then these individuals would be defiers that would invalidate the monotonicity assumption needed for correct inferences (Angrist et al., 1996). The hierarchies of selves approach proposes that a more robust method to infer norms is to measure the impact of a prime among those who not only belong to a social category, but also value it highly in relative and absolute terms.

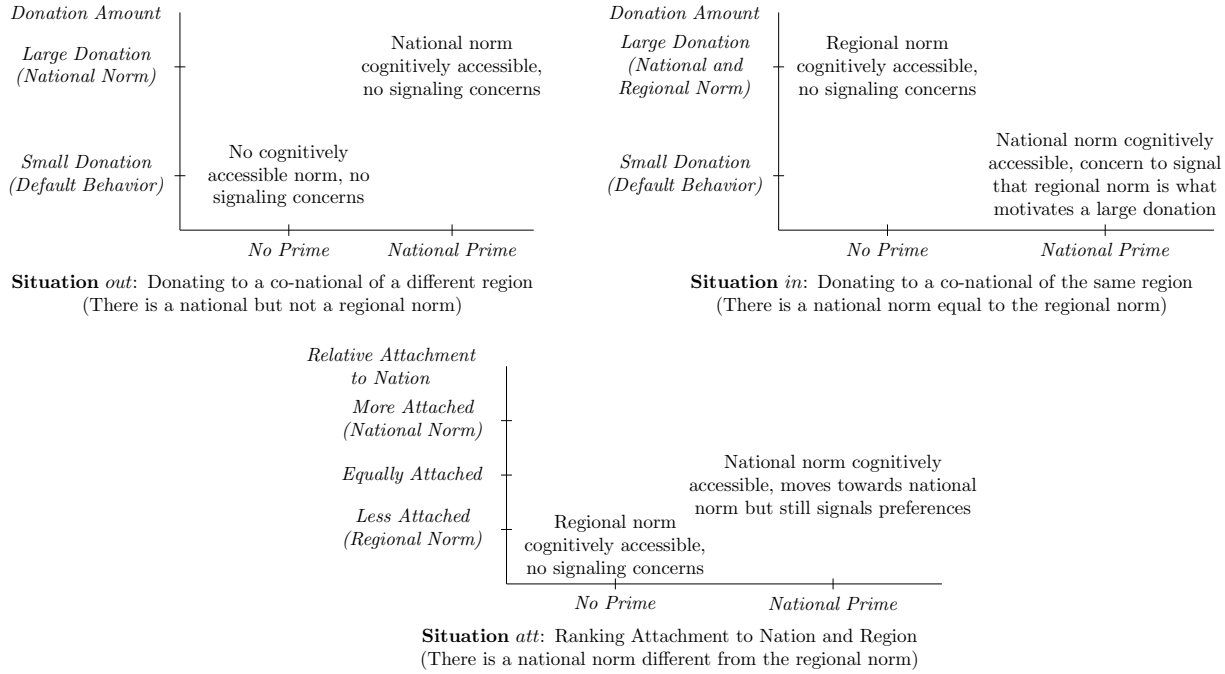
A Tables And Figures For Body Of Text

	Regional Norm	National Norm
Donating to someone of the same region	Large donation	Large donation
Donating to co-national of a different region	No norm	Large donation

Table 1: National and Regional Norms in Different Situations

	Hierarchies of selves	Framing	Signaling
Donating to someone of the same region	Away	Towards	Away
Donating to co-national of a different region	Towards	Towards	Away

Table 2: Does a National Prime Move Regionalists Away Or Towards The National Norm?



Situation *out*: Donating to a co-national of a different region
(There is a national but not a regional norm)

Situation *in*: Donating to a co-national of the same region
(There is a national norm equal to the regional norm)

Situation *att*: Ranking Attachment to Nation and Region
(There is a national norm different from the regional norm)

Figure 1: Graphical Represenation of Situations *in*, *out*, and *att*

Baseline Questions
State X subject is most attached to
Region Y subject believes corresponds to state X
Attachment to region and to nation
Other baseline covariates
Randomization
Randomization blocked on the difference between attachment to region Y and nation.
Assignment to one of four conditions: 1) Randomly chosen state from region Y without a prime 2) Randomly chosen state from region Y with a prime 3) Randomly chosen state outside of region Y without a prime 4) Randomly chosen state outside of region Y with a prime
In each condition, a rich or a poor state from the region was chosen with equal probability if possible.
Endline Questions *(Two Weeks After Baseline)*
Exposition to donation petition according to random assignment
Free word association and NGO name guess
Donation decision
Questions about NGO, state
Attachment to region Y and to nation

Table 3: Summary of Design

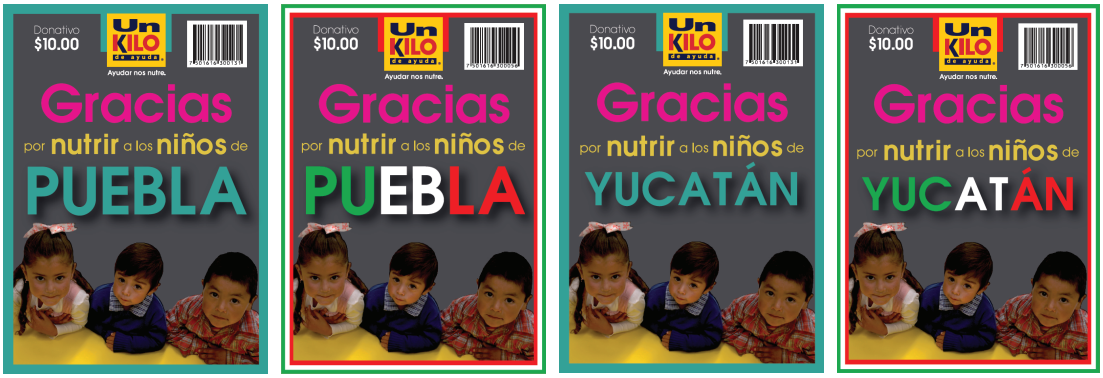


Figure 2: Treatment Cards. The first and third cards are real colored in the frame and letters of the state. The second and fourth replace the teal with the green white and red of the Mexican flag.

	Mean	S.D.	N
Male	0.59	0.49	580
Age 17 To 18	0.13	0.34	580
Age 19 To 20	0.37	0.48	580
Age 21 To 22	0.32	0.47	580
White	0.37	0.48	580
Trusts The NGO	0.45	0.50	580
Contributes To The NGO	0.51	0.50	580
Responded After First Round of Emails	0.56	0.50	580
Responded After Second Round of Emails	0.22	0.42	580
Self-Reported Tenth Income Decile	0.16	0.37	580
Self-Reported Ninth Income Decile	0.13	0.34	580
Self-Reported Eighth Income Decile	0.23	0.42	580
Self-Reported Seventh Income Decile	0.27	0.44	580
Self-Reported Sixth Income Decile	0.13	0.33	580
From Mexico City	0.49	0.50	580
From State of Mexico	0.09	0.28	580
From Nuevo León	0.04	0.20	580

Table 4: Summary Statistics of Selected Covariates

Regional Attachment	National Attachment 1	2	3	4	5	**Total**
1	4	5	7	5	3	24
2	2	7	24	26	10	69
3	0	13	44	72	40	169
4	1	6	23	86	103	219
5	1	1	12	31	61	106
Total	8	32	110	220	217	587

Table 5: Self-Reported Regional and National Attachment On A Five Point Scale. Regionalists are those more attached to the region, nationalists those more attached to the nation

	Mean	S.D.	N
Regionalist (More Attached To Region Than To Nation)	0.15	0.36	580
Nationalist (More Attached To Nation Than To Region)	0.50	0.50	580
Donation Petition Contained A Prime	0.51	0.50	580
Was Asked To Donate To An In-region State	0.49	0.50	580
Was Asked To Donate To A Poor State	0.40	0.49	580
Was Asked To Donate To A Relatively Poor State In The Region	0.50	0.50	508
Number Of States in Self-Defined In-Region	4.07	2.35	580
Donation Out Of 100 Pesos	63.59	41.15	571
Thinks Of Assigned State As Part Of A Region	0.42	0.49	571
Thinks Of Assigned State Like Others In Country	0.31	0.46	571
Regionalist In Endline Survey	0.18	0.39	555
Nationalist In Endline Survey	0.39	0.49	555

Table 6: Summary Statistics of Dependent and Independent Variables

	Regionalists	
	Not primed	Primed
In-region	81.4	50.7
	(11.78)	(13.64)
Out-region	49.2	74.4
	(11.86)	(10.15)

	Nationalists	
	Not primed	Primed
In-region	64.4	62.5
	(5.51)	(6.11)
Out-region	66.6	59.6
	(5.45)	(6.58)

Table 7: Average Donation Decisions by Region, Priming and Strength of Attachment. Robust standard errors in parentheses.

	(1) Donation	(2) Donation
In-Region	-10.96	-2.267
	(7.650)	(8.383)
Prime	-7.663	-7.045
	(6.891)	(8.014)
Regionalist	-19.20*	-17.45
	(10.50)	(12.85)
Prime*InRegion	3.454	5.142
	(10.80)	(12.05)
Regionalist*Prime	16.49	32.30**
	(15.09)	(15.93)
Regionalist*In-Region	17.44	34.49*
	(15.15)	(17.81)
Triple Interaction	-28.43	-61.12**
	(24.02)	(26.16)
Observations	372	372
R^2	0.150	0.396
Controls	No	Yes
Donor and Region FE	Yes	Yes

Robust standard errors in parentheses

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

Dependent variable of the two columns are donation decisions out of 100 Mexican pesos to a charitable cause benefiting a certain recipient state. The independent variables in the table describe whether the recipient state was in their self-defined sub-national region, whether the donation petition contained a prime, and whether they had reported being more attached to the region than to the nation in a baseline survey two weeks earlier. All columns include recipient state fixed effects, and fixed effects of the state of the donor. The second column includes additional controls: party preference dummies, income level dummies, trust in institution dummies, contribution to institution dummies, age dummies, low attachment to nation and region dummies, gender, race, baseline and endline location variables, baseline and endline round dummies for the round of response to the email invitation, and size of region variables.

Table 8: The Impact of Recipient Region, Priming and Relative Strength of Attachment On Donation Decisions

	(1) Donation	(2) Donation	(3) Donation	(4) Donation	(5) Donation	(6) Donation
In-region	-10.34**	-6.023			-9.235	-0.842
	(4.489)	(4.873)			(5.733)	(6.446)
Prime			-4.505	-1.998	-3.469	2.447
			(3.758)	(3.907)	(4.936)	(5.294)
Prime*In-region					-2.255	-9.121
					(7.395)	(7.944)
Constant	59.45***	166.7	48.64***	183.3	67.41***	189.2
	(16.99)	(118.0)	(17.27)	(115.9)	(18.63)	(115.8)
Observations	568	568	568	568	568	568
R^2	0.099	0.252	0.093	0.256	0.102	0.261
Controls	No	Yes	No	Yes	No	Yes
Donor and Region FE	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses

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

Dependent variable of the six columns are donation decisions out of 100 Mexican pesos to a charitable cause benefiting a certain recipient state. The independent variables in the table describe characteristics of the recipient state and whether the donation petition contained a national prime. All columns include recipient state fixed effects, and fixed effects of the state of the donor. The second, fourth and sixth columns include additional controls: party preference dummies, income level dummies, trust in institution dummies, contribution to institution dummies, age dummies, low attachment to nation and region dummies, gender, race, baseline and endline location variables, baseline and endline round dummies for the round of response to the email invitation, and size of region variables.

Table 9: The Impact of Region and Priming On Donation Decisions

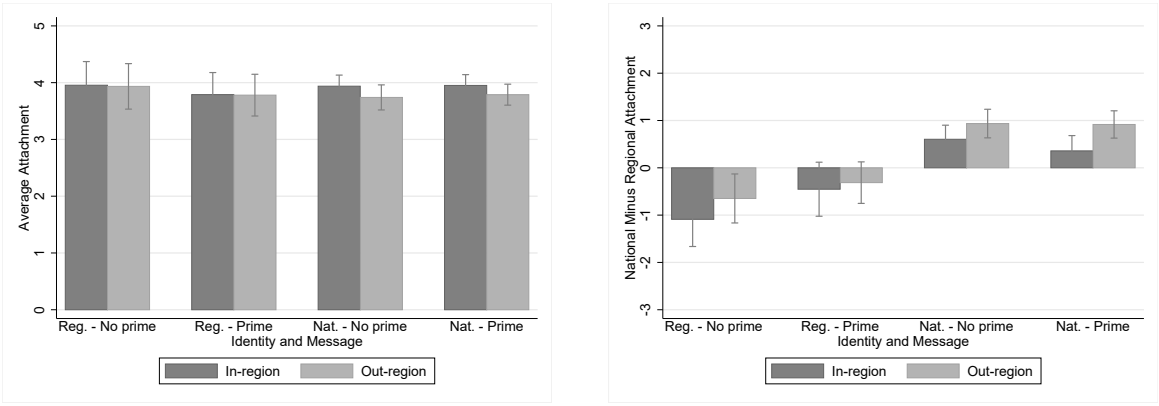


Figure 3: The Impact of Region, Priming and Relative Strength of Attachment on Two Measures: average endline regional and national attachment (left) and their difference (right)

	(1) Difference	(2) Difference
Prime	-0.116	-0.101
	(0.131)	(0.140)
Regionalist	-1.726***	-1.590***
	(0.196)	(0.250)
Regionalist*Prime	0.563**	0.592**
	(0.254)	(0.292)
Observations	361	361
Adjusted R^2	0.263	0.265
Controls	No	Yes
Donor and Region FE	Yes	Yes

Standard errors in parentheses

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

Dependent variable of the two columns are the difference between self-reported attachment to the nation and attachment to the region at the endline. The independent variables in the table describe whether the donation petition they were exposed to contained a national prime, and whether they had self-reported being more attached to the region than to the nation in a baseline survey at least two weeks before the endline. All columns include recipient state fixed effects, and fixed effects of the state of the donor. The second column includes additional controls: party preference dummies, income level dummies, trust in institution dummies, contribution to institution dummies, age dummies, low attachment to nation and region dummies, gender, race, baseline and endline location variables, baseline and endline round dummies for the round of response to the email invitation, and size of region variables.

Table 10: Difference Between National and Regional Attachment at Endline

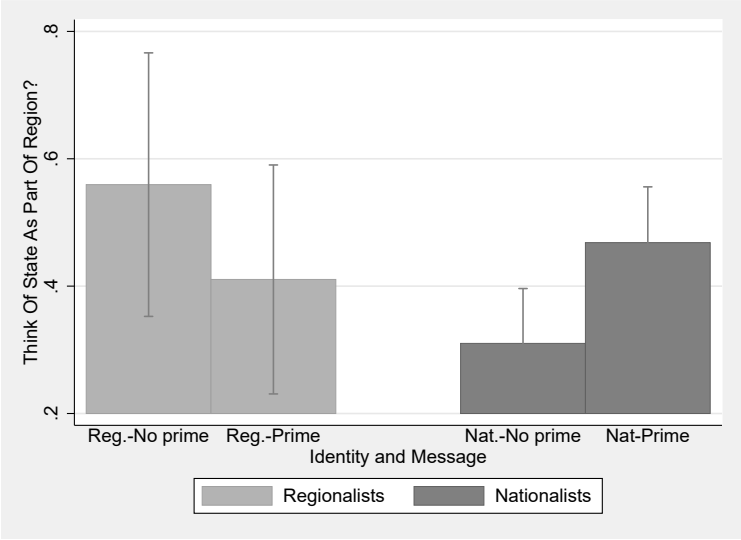


Figure 4: The Impact of Relative Strength Of Attachment And Priming: 'Is it easier for you to think of the recipient state as part of the nation or as part of a region?'

	(1) Thought Of As Region	(2) Thought Of As Region
Regionalist	0.220**	0.249**
	(0.100)	(0.120)
Prime	0.106*	0.158**
	(0.0625)	(0.0630)
Prime*Regionalist	-0.288**	-0.307**
	(0.137)	(0.141)
Constant	1.205***	-1.785
	(0.403)	(1.935)
Observations	372	372
Adjusted R^2	0.012	0.069
Constant	No	Yes
Donor and Region FE	Yes	Yes

Standard errors in parentheses

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

Dependent variable of the two columns are the answer to the question 'Is it easier for you to think of the donation state as part of the nation or part of a region?' The independent variables in the table describe whether the donation petition they were exposed to contained a national prime, and whether they had self-reported being more attached to the region than to the nation in a baseline survey at least two weeks before the endline. All columns include recipient state fixed effects, and fixed effects of the state of the donor. The second column includes additional controls: party preference dummies, income level dummies, trust in institution dummies, contribution to institution dummies, age dummies, low attachment to nation and region dummies, gender, race, baseline and endline location variables, baseline and endline round dummies for the round of response to the email invitation, and size of region variables.

Table 11: Prime Affects How States Are Thought Of

B Proof of Proposition 1

Proof. We begin by characterizing the equilibria for the utility function u_{BT}.

Suppose individuals are pooling on $d \neq n(s, N)$, and consider a deviation to $n(s, N)$. The utility for players of type θ with u_{BT} is:

$$h_s\mu - (d - n(s, \theta))^2 > h_s\hat{\mu} - (n(s, N) - n(s, \theta))^2$$

The D1 criterion asks us to look for the set of strategies the audience would follow such that each type will want to deviate. Player of type θ will do better off deviating for all out-of-equilibrium beliefs:

$$\hat{\mu}_\theta \geq \mu + \frac{-(d - n(s, \theta))^2 + (n(s, N) - n(s, \theta))^2}{h_s}$$

It is easy to verify that for all $d \in [\min D, \max D)$ with $\max D = n(s, N)$ and $n(s, R) - \min D < n(s, N) - n(s, R)$, $\hat{\mu}_R > \hat{\mu}_N = 0$. Therefore, the set of audience strategies that makes N want to deviate to $n(s, N)$ is larger than the set for R types. By the D1 criterion, out-of-equilibrium beliefs when choosing $n(s, N)$ must be at least equal to μ. But then nationalists would want to deviate to $n(s, N)$.

I will now show that the separating equilibrium with $d^*(s, N, N, h_s) = n(s, N)$ and $d^*(s, N, R) = n(s, R)$ can be sustained. Nationalists are obtaining their maximum utility. Therefore, any out-of-equilibrium decision can be reasonably justified with the belief that the decision was taken by a regionalist. This implies that regionalists then prefer $n(s, R)$ to any decision other than $n(s, N)$, and they prefer $n(s, R)$ to $n(s, N)$ if and only if:

$$0 \geq hs - (n(s, N) - n(s, R))^2 \tag{1}$$

If there is a type-dependent strategy in which types separate and regionalists do not choose $n(s, R)$, regionalists always do better off by deviating to $n(s, R)$. If there is a type-dependent strategy where regionalists choose $n(s, R)$, nationalists do not choose $n(s, N)$ and (1) holds, then the only reasonable beliefs are for only nationalists to choose $n(s, N)$, so they would find it optimal to do so.

Finally, suppose players are pooling on $n(s,N)$. Under the assumptions, any other choice d will be closer to the regional norm and farther from the national norm. Therefore, for any out-of-equilibrium belief, regionalists will be doing relatively better off than nationalists with d. By the D1 criterion, we can then have reasonable out-of-equilibrium beliefs that d is chosen by regionalists. But then if (1) holds, neither regionalists nor nationalists would want to deviate.

I have established that there are two possible equilibria with u_{BT}. It is easy to verify that with u_S, there are also two equilibria: one where all pool on $n(s,N)$, and one where nationalists chooses $n(s,N)$ and they separate. Nationalists always choose $n(s,N)$. Regionalists choose $n(s,N)$ in equilibrium if and only if:

$$h_s(1-\mu) \geq h_s - (d^c - n(s,N))^2$$

where $d^c \equiv \arg\max_{d \in D \setminus n(s,N)} -(n(s,N)-d)^2$ is the closest choice to $n(s,N)$. Otherwise, they choose d_c. $\square$

C Proof of Result 2

If $n(s,N) = n(s,R)$, $d^*(s,R,\theta) = n(s,N)$ by Lemma 1 and $d^*(s,R,N) \neq n(s,N)$ by Lemma 2.

If $0 = n(s,R) \neq n(s,N)$, then $d^*(s,R,\theta) = n(s,R)$ by Lemma 1 and $d^*(s,R,N) = n(s,N)$ by Lemma 2 considering $h_s = 0$.

If $\exists\ d \in (n(s,R), n(s,N))$ with $0 \neq n(s,R)$, $d^*(s,R,\theta) = n(s,R) \neq \{d^c, n(s,N)\}$ by Lemma 1 and $d^*(s,R,N) = d^c \neq n(s,N)$ by Lemma 2 considering $h_s \geq h^*$.

If $\nexists d \in (n(s,R), n(s,N))$ with $0 \neq n(s,R) \neq n(s,N)$, $d^*(s,R,\theta) = n(s,R)$ by Lemma 1 and $d^*(s,R,N) = n(s,R)$ by Lemma 2 considering $h_s \geq h^*$.

The other direction comes from the fact that I have exhausted all possibilities. $\square$

D Balance Tables

Tables 12, 13 and 14 present the balance over recipient state location, primes and regionalism separately. In general, the treatments are balanced. The differences that were unexpected are small and few. Those who receive in-region cards were a bit more likely to state that they had

given to the NGO in the baseline (Table 12). Those who receive a prime were a bit more likely to have a recorded IP address from the college (Table 13). Regionalists trust the government slightly more (Table 14).

The expected differences are large.

In-region versus out-region states (Table 12): In-region states are significantly less poor than out-region states. This is driven by the fact that half of the sample is from Mexico City, which is from a relatively rich region of the country. This can be seen in that there are more in-region cards going to Mexico City than out-region cards.

As we've discussed, I maximized balance in recipient state income across treatments by randomizing whether subjects were assigned to a relatively rich or a relatively poor state in the region they were assigned to. When the region did not have one of the poorest states, a middle income state was assigned. As I mentioned previously, the label 'Poorer' is a dummy for whether the subject was assigned a relatively poor state. It restricts the sample to those who had a relatively poor and a relatively rich state in the region they were assigned, which are 508 out of the 580 observations. This variable is well balanced across all the treatments.

The variable 'Poorer plus' includes states with only rich states in their region, defining them as relatively richer states and states with only poor states in their region, defining them as relatively poor states. This accounts for all but 4 of the observations in the data. This variable is balanced.

Regionalists versus nationalists (Table 14): The overall balance between regionalists and nationalists is particularly important given that attachment is not randomized. I show the results of comparing regionalists to 'weak' nationalists, or those who value their national identity at least as much as their regional identity. The same results hold if I compare regionalists to 'strong' nationalists, or those who are more attached to the nation than to the region.

There are three significant differences between regionalists and weak nationalists.

1. *Attachment:* Mechanically, regionalists are significantly more attached to their region and significantly less attached to their nation than weak nationalists.

2. *Region size:* Regionalists included more states when defining their region than weak nationalists. It is intuitive that regionalists define their region differently than weak nationalists do, although one may be concerned that it is region size that is driving the results. This

	Out-Region		In-Region		Mean diff
	Mean	S.D	Mean	S.D.	t
Male	0.57	(0.50)	0.61	(0.49)	-0.96
Age 17 To 18	0.13	(0.33)	0.14	(0.35)	-0.58
White	0.36	(0.48)	0.40	(0.49)	-1.01
Trusts NGO	0.45	(0.50)	0.45	(0.50)	0.21
Gives to NGO	0.47	(0.50)	0.55	(0.50)	-2.05*
Trusts government	3.41	(1.94)	3.48	(2.05)	-0.38
Gives to street children	3.39	(1.87)	3.52	(1.84)	-0.88
Latitude	0.02	(0.15)	0.03	(0.17)	-0.40
First Round	0.53	(0.50)	0.58	(0.49)	-1.30
First Round Final	0.46	(0.50)	0.48	(0.50)	-0.48
Votes PAN	0.58	(0.49)	0.60	(0.49)	-0.62
ITAM IP	0.30	(0.46)	0.30	(0.46)	-0.02
High Income	0.16	(0.37)	0.16	(0.37)	-0.04
Poor	0.49	(0.50)	0.29	(0.46)	5.07***
Poorer	0.49	(0.50)	0.49	(0.50)	-0.13
Poorer plus	0.51	(0.50)	0.56	(0.50)	-1.38
Own State Poor	0.82	(0.38)	0.83	(0.37)	-0.37
From Mexico City	0.50	(0.50)	0.48	(0.50)	0.56
From State of Mexico	0.08	(0.27)	0.10	(0.30)	-0.85
From Nuevo León	0.04	(0.20)	0.04	(0.20)	0.02
To Mexico City	0.04	(0.19)	0.25	(0.43)	-7.68***
States in Region	3.95	(2.26)	4.20	(2.44)	-1.26
Regionalist	0.15	(0.36)	0.16	(0.36)	-0.14
Nationalist	0.50	(0.50)	0.50	(0.50)	0.04

Table 12: Balance Table for State Location

seems implausible. Individuals are randomly assigned only one state within their region, and in-region and out-region states are balanced on income. Further, we will include recipient state fixed effects in the analysis. Therefore, what will matter in the test will not depend on the size of the region or be the region-specific characteristics, but on whether the assigned state belongs to the individuals' region.

3. *Trust in government:* Regionalists are more likely than weak nationalists to trust the government. Given the amount of variables considered and the slight significance of the difference, this may be due to random chance. Further, there again seems to be a lack of an obvious causal chain that would drive the results.

References

Afridi, F., S. X. Li, and Y. Ren (2015). Social identity and inequality: The impact of china's hukou system. *Journal of Public Economics 123*, 17–29.

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *Quarterly journal of Economics*, 715–753.

	No Prime		Prime		Mean diff
	Mean	S.D.	Mean	S.D.	t
Male	0.62	(0.49)	0.56	(0.50)	1.38
Age 17 to 18	0.14	(0.35)	0.12	(0.33)	0.71
White	0.40	(0.49)	0.35	(0.48)	1.33
Trust NGO	0.43	(0.50)	0.47	(0.50)	-1.20
Gives to NGO	0.52	(0.50)	0.50	(0.50)	0.45
Regionalist	0.16	(0.36)	0.15	(0.36)	0.21
Nationalist	0.50	(0.50)	0.51	(0.50)	-0.12
Trust in Government	3.33	(1.92)	3.56	(2.06)	-1.44
Gives to street children	3.49	(1.92)	3.42	(1.79)	0.48
Latitude	0.03	(0.18)	0.02	(0.13)	1.30
First round	0.56	(0.50)	0.55	(0.50)	0.12
First round final	0.46	(0.50)	0.47	(0.50)	-0.37
Votes PAN	0.61	(0.49)	0.57	(0.50)	1.04
ITAM IP	0.34	(0.47)	0.26	(0.44)	2.15*
High income	0.17	(0.37)	0.15	(0.36)	0.55
Poor	0.39	(0.49)	0.40	(0.49)	-0.12
Poorer	0.49	(0.50)	0.48	(0.50)	0.22
Poorer plus	0.53	(0.50)	0.54	(0.50)	-0.37
Own State Poor	0.84	(0.37)	0.82	(0.39)	0.78
From Mexico City	0.50	(0.50)	0.48	(0.50)	0.54
From State of Mexico	0.10	(0.30)	0.08	(0.27)	0.86
From Nuevo León	0.04	(0.21)	0.04	(0.20)	0.20
To Mexico City	0.15	(0.36)	0.13	(0.34)	0.70
States in region	4.01	(2.38)	4.13	(2.32)	-0.60

Table 13: Balance Table for Primes

	Weak Nationalist		Regionalist		Mean diff
	Mean	S.D.	Mean	S.D.	t
Male	0.58	(0.49)	0.67	(0.47)	-1.63
Age 17 To 18	0.13	(0.33)	0.17	(0.37)	-0.94
White	0.38	(0.49)	0.36	(0.48)	0.41
Trust in NGO	0.46	(0.50)	0.41	(0.49)	0.80
Gives to NGO	0.51	(0.50)	0.54	(0.50)	-0.69
Trusts government	3.37	(1.98)	3.87	(2.00)	-2.18*
Gives to street children	3.46	(1.84)	3.42	(1.91)	0.17
Latitude	0.02	(0.15)	0.03	(0.18)	-0.45
First round	0.56	(0.50)	0.52	(0.50)	0.68
First round final	0.46	(0.50)	0.48	(0.50)	-0.23
Votes PAN	0.60	(0.49)	0.57	(0.50)	0.51
ITAM IP	0.31	(0.46)	0.24	(0.43)	1.23
High Income	0.16	(0.37)	0.14	(0.35)	0.41
Poor	0.39	(0.49)	0.42	(0.50)	-0.53
Poorer	0.49	(0.50)	0.50	(0.50)	-0.20
Poorer plus	0.54	(0.50)	0.52	(0.50)	0.35
Own state poor	0.83	(0.38)	0.81	(0.39)	0.44
From Mexico City	0.50	(0.50)	0.41	(0.49)	1.62
From State of Mexico	0.09	(0.29)	0.08	(0.27)	0.41
From Nuevo León	0.04	(0.19)	0.07	(0.25)	-1.02
To Mexico City	0.14	(0.35)	0.14	(0.35)	-0.14
States in region	4.26	(2.40)	3.06	(1.74)	5.65***
Regional attachment	3.39	(1.02)	4.31	(0.80)	-9.53***
National Attachment	4.21	(0.84)	3.03	(0.87)	11.93***

Table 14: Balance Table for Regionalism

Akerlof, G. A. and R. E. Kranton (2005). Identity and the economics of organizations. *The Journal of Economic Perspectives 19*(1), 9–32.

Akerlof, G. A. and R. E. Kranton (2010). Identity economics: How identities shape our work, wages, and well-being.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Bénabou, R. and J. Tirole (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics 126*(2), 805–855.

Benjamin, D., J. Choi, et al. (2010). Social identity and preferences. *American Economic Review 100*(4), 1913–28.

Benjamin, D. J., J. J. Choi, and G. Fisher (2016). Religious identity and economic behavior. *Review of Economics and Statistics 98*(4), 617–637.

Berger, J., M. Meredith, and S. C. Wheeler (2008). Contextual priming: Where people vote affects how they vote. *Proceedings of the National Academy of Sciences 105*(26), 8846–8849.

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman (2010). What's advertising content worth? evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics 125*(1), 263–306.

Bhattacharjee, A., J. Berger, and G. Menon (2014). When identity marketing backfires: consumer agency in identity expression. *Journal of Consumer Research 41*(2), 294–309.

Boschini, A., A. Dreber, E. von Essen, A. Muren, E. Ranehill, et al. (2014). Gender and economic preferences in a large random sample. Technical report, Stockholm University, Department of Economics.

Cadsby, C. B., M. Servátka, and F. Song (2013). How competitive are female professionals? a tale of identity conflict. *Journal of Economic Behavior & Organization 92*, 284–303.

Chang, D., R. Chen, and E. Krupka (2015). Social norms and identity dependent preferences. *Univ Mich Work Pap*.

Charnysh, V., C. Lucas, and P. Singh (2014). The ties that bind national identity salience and pro-social behavior toward the ethnic other. *Comparative Political Studies*, 0010414014543103.

Chen, Y. and S. X. Li (2009). Group identity and social preferences. *The American Economic Review 99*(1), 431–457.

Chen, Y., S. X. Li, T. X. Liu, and M. Shih (2014). Which hat to wear? impact of natural identities on coordination and cooperation. *Games and Economic Behavior 84*, 58–86.

Cohn, A., E. Fehr, and M. A. Maréchal (2014). Business culture and dishonesty in the banking industry. *Nature 516*(7529), 86–89.

Cohn, A., E. Fehr, and M. A. Maréchal (2015). A culture of gambling? evidence from the banking industry. Technical report, Working Paper.

Cohn, A. and M. A. Maréchal (2016). Priming in economics. *Current Opinion in Psychology 12*, 17–21.

Cohn, A., M. A. Maréchal, and T. Noll (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies 82*(4), 1289–1308.

Dee, T. S. (2014). Stereotype threat and the student-athlete. *Economic Inquiry 52*(1), 173–182.

Fryer, R. G., S. D. Levitt, and J. A. List (2008). Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study. *The American Economic Review 98*(2), 370–375.

Gaertner, S. L. and J. F. Dovidio (2012). Common ingroup identity model. *The encyclopedia of peace psychology*.

Gilbert, D., G. King, S. Pettigrew, and T. Wilson (2016). More on estimating the reproducibility of psychological science. *Available at projects. iq. harvard. edu/files/psychology-replications/files/gkpw_post_publication_response. pdf*.

Gómez, Á., J. F. Dovidio, C. Huici, S. L. Gaertner, and I. Cuadrado (2008). The other side of we: When outgroup members express common identity. *Personality and Social Psychology Bulletin*.

Guala, F. and A. Filippin (2016). The effect of group identity on distributive choice: Social preference or heuristic? *The Economic Journal*.

Hoff, K. and P. Pandey (2014). Making up peoplethe effect of identity on performance in a modernizing society. *Journal of Development Economics 106*, 118–131.

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, 0959354314529616.

Krupka, E. L. and R. A. Weber (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association 11*(3), 495–524.

LeBoeuf, R. A., E. Shafir, and J. B. Bayuk (2010). The conflicting choices of alternating selves. *Organizational Behavior and Human Decision Processes 111*(1), 48–61.

Ledgerwood, A. and S. Chaiken (2007). Priming us and them: automatic assimilation and contrast in group attitudes. *Journal of personality and social psychology 93*(6), 940.

Liverman, D. and A. Cravey (1992). Geographic perspectives on mexican regions. *Mexico's Regions: Comparative History and Development*, 39–57.

Martin, D. S., T. R. Rogers, et al. (1984). Australian nationalism and mass media persuasian appeals. *Media Information Australia* (32), 33.

McKay, R., C. Efferson, H. Whitehouse, and E. Fehr (2011). Wrath of god: religious primes and punishment. *Proceedings of the Royal Society of London B: Biological Sciences 278*(1713), 1858–1863.

Morris, M. W., E. Carranza, and C. R. Fox (2008). Mistaken identity activating conservative political identities induces conservative financial decisions. *Psychological Science 19*(11), 1154–1160.

Pedic, F. (1990). Persuasiveness of nationalistic advertisements. *Journal of Applied Social Psychology 20*(9), 724–738.

Puntoni, S., S. Sweldens, and N. T. Tavassoli (2011). Gender identity salience and perceived vulnerability to breast cancer. *Journal of Marketing Research 48*(3), 413–424.

Reed, A. (2004). Activating the self-importance of consumer selves: Exploring identity salience effects on judgments. *Journal of consumer research 31*(2), 286–295.

Sachs, N. (2009). Experimenting with identity: Islam, nationalism and ethnicity. In *APSA 2009 Toronto Meeting Paper*.

Seiter, J. S. and R. H. Gass (2005). The effect of patriotic messages on restaurant tipping1. *Journal of Applied Social Psychology 35*(6), 1197–1205.

Shayo, M. (2009). A model of social identity with an application to political economy: Nation, class, and redistribution. *American Political science review 103*(02), 147–174.

Shih, M., T. L. Pittinsky, and N. Ambady (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological science 10*(1), 80–83.

Swann, W. B. and J. K. Bosson (2010). Self and identity. *Handbook of social psychology*.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. CUP Archive.

Tajfel, H. and J. C. Turner (1979). An integrative theory of intergroup conflict. *The social psychology of intergroup relations 33*(47), 74.

Thomas, T. C., R. K. Trump, and L. L. Price (2015). Advertising as unfavorable self-presentation: The dirty laundry effect. *Journal of Advertising 44*(1), 58–70.

Transue, J. E. (2007). Identity salience, identity acceptance, and racial policy attitudes: American national identity as a uniting force. *American Journal of Political Science 51*(1), 78–91.

Turner, J. C. and K. J. Reynolds (2012). Self categorization theory. In P. A. M. V. Lange, A. W. Kruglanski, and E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology*, Volume 2, Chapter 46, pp. 379–398. London: SAGE Publications Ltd.

Wichardt, P. C. (2008). Identity and why we cooperate with those we do. *Journal of Economic Psychology 29*(2), 127–139.

E Surveys In English And In Spanish