Evaluating the Influence of International Norms and Shaming on State Respect for Rights: An Audit Experiment with Foreign Embassies

Zhanna Terechshenko *
Pennsylvania State University
zvt5048@psu.edu

Charles Crabtree
University of Michigan
ccrabtr@umich.edu

Kristine Eck
Uppsala University
Kristine.Eck@pcr.uu.se

Christopher J. Fariss
University of Michigan
cjf0006@gmail.com

October 28, 2018

Abstract

How do international norms affect respect for human rights? In this research note, we report the results of an audit experiment with foreign missions that investigates the extent to which state agents observe international norms and react to the potential of international shaming. Our experiment involved emailing 669 foreign diplomatic missions in the United States, Canada, and the United Kingdom with requests to contact domestic prisoners. According to the United Nations, prisoners have the right for individuals to contact them. We randomly varied (1) whether we reminded embassies about the existence of an international norm permitting prisoner contact and (2) whether the putative email sender is associated with a fictitious human rights organization and, thereby, has the capacity to shame missions through naming and shaming for violating this norm. We find strong evidence for the positive effect of international norms on state respect for human rights. Contra to expectation, though, we find that the potential of international shaming *does not* increase the probability of state compliance. Indeed, the positive effect of the norms cue disappears when it is coupled with the shaming cue, suggesting that shaming might have a 'backfire' effect.

*We would like to thank Rob Carrol, Eric Gartzke, Nadiya Kostyuk, and Barry Hashimoto for their extremely helpful comments. We would also like to thank attendees of the 2016 Peace Science Annual Meeting and the 2017 meetings of the American Political Science Association and the Midwest Political Science Association. The Pennsylvania Institutional Review Board approved this study (IRB STUDY00004057).

1 Introduction

How do international norms affect state respect for human rights? The theoretical literature argues that states' compliance with human rights norms is induced by a socialization process through which compliance is encouraged by "diplomatic praise or censure . . . which is reinforced by material sanctions and incentives" (Finnemore and Sikkink, 1998). After a sufficient number of states are persuaded by norm entrepreneurs to adopt a new international norm, other states experience an effect similar to "peer pressure," which induces them to adhere to this new norm. One possible explanation for the effectiveness of this mechanism is that states want to demonstrate that they are legitimate members of the international human rights regime. Another possible explanation is that leaders want to make a good international impression in order to "enhance national esteem"(Creamer and Simmons, 2015; Finnemore and Sikkink, 1998; Hill, 2010).

Perhaps the most convincing explanation, though, is that states abide by international norms to placate other international actors who also act as agents of socialization (Finnemore and Sikkink, 1998). This is particularly true in the case of human rights organizations (HRO), who can 'name and shame' regimes that do not comply with international standards. Although HROs have limited power to directly sanction regimes, they can impose reputation costs on a regime by shaming them and thereby help spur public condemnation that might lead to the imposition of political and/or economic sanctions, such as decreased multilateral aid allocations (Lebovic and Voeten, 2009; Krain, 2012)[1]. This suggests that while other factors might influence state compliance with international norms, this compliance should be higher in the presence of HRO oversight. In line with this logic, 'naming and shaming' campaigns are considered to be one of the best strategies for enforcing human rights norms (Brysk, 1993; Keck and Sikkink, 1998; Meernik et al., 2012; Murdie and Davis, 2012).

While these theoretical claims about the influence of international norms and shaming

[1]HROs can also increase social grievances and legitimize regime opponents (Brysk, 1993; Keck and Sikkink, 1998)

are intuitively appealing, the empirical support for these effects is less clear. For example, though a large quantitative literature has focused on understanding the effect of human rights treaty ratification on states' respect for physical integrity rights, the nature of this relationship continues to remain contested (Creamer and Simmons, 2015; Fariss, 2014). One problem is that many studies rely on observational research designs, from which it is often difficult to make causal claims (Winship and Morgan, 1999).[2] Moreover, the effect of norms and shaming on some rights—such as empowerment rights or prisoners' rights—remains almost wholly unstudied. We address these issues by focusing on prisoner's rights leveraging an experimental design. We also re-focus the analysis by looking at the state agents who implement policy. The literatures norms and shaming have focused on state-level analyses of how the executive or judiciary respond to international norms or shaming. Less consideration has been given to whether state agents respect human rights in practice. This is potentially an important oversight, as individual agents of the state have considerable agency to influence human rights practices (Lipsky, 2010; Wilson, 1989); frequent media reports of police brutality and disrespect for equality rights are indicative of the centrality of bureaucratic compliance. There is a growing appreciation that studying state agents can facilitate a more robust understanding of human rights practices.

To address these gaps in the literature conducted a field experiment with foreign embassies to test the extent to which states observe international norms and react to the potential of international shaming. Specifically, we examine (1) the effect of international norms on respect for prisoners' rights and (2) the effect of potential shaming. We did this by emailing foreign diplomatic missions in the United States, Canada, and the United Kingdom with requests to contact domestic prisoners, a right acknowledged by the United Nations. To test the effect of norms, we randomly varied whether we reminded embassies about the existence of an international norm permitting prisoner contact. To test the effect of possible shaming, we randomly varied whether the putative email sender is associated with a fic-

[2]The primary issue here is that without a credible identification strategy, researchers cannot know if the estimated effect of norms or shaming is unconfounded (Pearl, 2009; Morgan and Winship, 2014).

tional human rights organization (HRO). The benefit of our research design is that it allows us to identify the causal effect of norms and shaming, while also examining how state agents respond to these factors.

Our results suggest that international norms have a strong effect on respect for human rights. Contra to expectation, though, we find that the potential of international shaming *does not* increase the probability of state compliance. Indeed, the positive effect of the norms cue disappears when it is coupled with the shaming cue, suggesting that shaming might have a 'backfire' effect.

This research note makes several contributions to the human rights literature. First, we situate prisoner's right within existing conceptualizations of human rights. Second, we focus on agents of the state in understanding human rights practices. Third, we provide empirical evidence for the effect of international norms on the behavior of state agents. This evidence suggests that international institutions can influence respect for human rights by creating self-enforcing norms. Fourth, we find that efforts to induce human rights compliance may have unintended and deliterious consequences. These contributions have clear policy implications for HROs and states seeking to influence respect for human rights.

2 Theory and Hypotheses

There are several ways to think about international human rights law. Some scholars argue that international law affects states' behavior through obligation, rule-making, and delegation (Abbott et al., 2000). Critics of this 'naive legalism' apply a regulative model to international law, assessing law's effectiveness in terms of its ability to regulate, constrain, or directly alter the behaviors of state leaders (Goldsmith and Krasner, 2003). According to the regulative model, the absence of centralized authority significantly undermines the coercive power of international law, thus reducing the probability of states' compliance and positive change in global human rights practices.

Recent improvements in human rights and the development of human rights law suggest,

however, that the most promising way of thinking about international law is the constitutive model. According to this model, international law is constituted by and, generative of, political struggles between the powerful (state) and the weak (rights claimants) (Dancy and Fariss, 2017). For instance, The International Campaign to Ban Landmines culminated in over 120 states signing the Ottawa Convention banning the use, stockpiling, production and transfer of antipersonnel landmines (Wexler, 2003). Another example of this process is the active work of women's rights networks, which fostered the adoption of a Declaration on the Elimination of Violence against Women by the United Nations General Assembly and the Inter-American Convention on the Prevention, Punishment, and Eradication of Violence against Women by the Organization of American States by 1994 (Keck and Sikkink, 1998).

The constitutive model assumes that international law does not need centralized enforcement (Dancy and Fariss, 2017). The proponents of this model emphasize the jurisgenerative power of international human rights norms to serve as a local framework for these political interactions (Benhabib, 2009; Dancy and Fariss, 2017). These political struggles occur not only within the states but outside of sovereign boundaries as well. Adoption of a norm by a large number of states redefines the appropriate behavior of other states as members of the international community in the process of socialization. In this context, socialization can be considered as a "mechanism through which norms leaders persuade others to adhere" (Finnemore and Sikkink, 1998, 902). States' adherence to international norms is one of the sources of international legitimation, which in turn contributes to the government's domestic basis of legitimation. In other words, states' compliance with international norms affects the compliance of their own citizens with government rules and law (Finnemore and Sikkink, 1998; Buchanan, 2007).

In addition, human rights law is a basis for a number of transnational social movements and advocacy groups or HROs, which also act as agents of socialization (Finnemore and Sikkink, 1998; Keck and Sikkink, 1998). These groups promote human rights norms implementation by pressuring target states and monitoring their compliance. For instance,

in order to alter the behavior of their targets, transnational advocacy groups use a variety of tactics, including information politics, symbolic politics, leverage politics, and accountability politics. Information politics refers to HROs' ability to generate politically usable information and mobilize it around their policy targets, while symbolic politics is the use of symbolic interpretation by advocacy groups in order to raise awareness about norms violation. Leverage politics includes increasing influence of HROs through seeking the leverage over more powerful actors. Finally, accountability politics is an effort of HROs to hold powerful actors accountable to their pronouncements, to the law, and to their contracts related to human rights. The effectiveness of these tactics varies with the issue under consideration, with the power of a HRO, and the target, whose behavior this HRO seeks to change (Keck and Sikkink, 1998).

All of these tactics are used in 'naming and shaming', which is perhaps the most effective method in human rights norms enforcement (Brysk, 1993; Keck and Sikkink, 1998; Meernik et al., 2012; Murdie and Davis, 2012). There is evidence that 'naming and shaming' campaigns by Amnesty International and the United Nations Commission on Human Rights (UNCHR) have significantly decreased the severity of ongoing instances of genocide and politicide from 1976 to 2008 (Krain, 2012). These campaigns include publicizing rights violations and pressuring perpetrators of rights violations to change their behavior (Krain, 2012). They can induce policy change through several mechanisms. Shaming reports are likely to have a significant impact on public opinion of human rights conditions, spurring policy change from the bottom-up (Ausderan, 2014; Davis, Murdie and Steinmetz, 2012; Dietrich and Murdie, 2015). HROs can also generate incentives for aid officials and institutions to change aid policy (Dietrich and Murdie, 2015). For example, 'shaming' in the UNCHR through resolutions that explicitly criticize governments for their human rights record tend to result in the reduction of multilateral aid received by the targets of these resolutions (Lebovic and Voeten, 2009).

The question that arises then is when do norms and 'naming and shaming' campaigns

affect state respect for human rights? For example, according to Murdie and Davis (2012), there two conditions under which shaming works. These are domestic presence of HROs within the targeted state and/or pressure by third-party actors. Terman and Voeten (2017) argue that the effect of naming and shaming is conditioned by the relationship between a state condemning human rights violations and the targeted state. Hafner-Burton (2008, p. 694) also emphasize selectivity of human rights organizations in publicizing abuses arguing that the effect of naming and shaming depends on "when and where the spotlight is shone". They also find that governments put under the spotlight are more likely to improve their records for political rights but not for political terror, and that incidents of terror are more likely to increase after the publication of the abuses. Hafner-Burton (2008) explain their findings by the fact that some governments might have limited capacity to protect certain rights or they might choose to improve some practices in order to increase their legitimacy and please the international community while continuing to violate other rights.

These studies, however, provide us with little or no information about the causality of the relationship between norms and shaming on human right practices. They also focus on a state-level analysis, largely neglecting the role of state agents in the human rights practice. Nevertheless, we do expect the effect of norms on state behavior to manifest in the decisions of individuals when they act as state representatives in bureaucratic entities. For instance, Lipsky (1980) argues that the attitudes and behaviors of street-level bureaucrats are affected by norms, which are usually derived from the society within which the bureaucratic agents live and work. Marrow (2009), citing Lewis and Ramakrishnan (2007), makes a similar point: "Likewise, some of the inclusive, service-oriented professional norms that affect bureaucrats behaviors today—such as ideals of pluralism and diversity in schools and the ideal of community policing in law enforcement—have grown out of past electoral political pressures."

In the process of professionalization, transnational advocacy networks influence individuals either directly or indirectly through the leaders of their political regime. In particular,

they do this by promoting professional norms and shared values and by providing technical assistance. As a result, states are subjected to pressure from international authorities and are encouraged to adopt and respect international norms in order to benefit from their membership in the international community. Thus, naming and shaming campaigns change leader and regime behavior by affecting their incentives, including reputation, domestic and/or international legitimacy, and foreign aid or foreign direct investment (Bush, 2011). Bureaucratic agents are central to the implementation of new policies adopted by leaders (Meier and O'Toole, 2006). Due to professional incentives, bureaucratic agents should be aware of human rights related norms and therefore be receptive to their usage (Lipsky, 1980; Lewis and Ramakrishnan, 2007).[3] This leads us to our *International Norms Hypothesis*.

> **International Norms Hypothesis**: State agents will exhibit greater compliance when reminded about international norms.

While HROs do not have enforcement power over rights perpetrators, they have the ability to shame them by putting these states and their violations in the international spotlight. They also can raise their concerns with more powerful actors, such as international organizations, that have leverage over states and affect their behavior. For example, Murdie and Peksen (2013) show that HRO activities increase the likelihood of economic sanctions against repressive regimes through information production and local population empowerment. Following Lewis and Ramakrishnan's (2007), we would expect bureaucratic agents who work at the embassies to be more receptive to the human rights related requests when there is a possibility of punishment. This expectation is captured in our *Shaming Hypothesis*.

> **Shaming Hypothesis**: State agents will provide more information as the requester's ability to shame the state increases.

In line with the literature, we also expect that the effect of international norms and shaming are reinforcing. This means that the positive effect of norms on compliance should

[3]This might be particularly true in the context of our experimental intervention, as bureaucrats who work at embassies likely have an increased understanding of international pressures and incentives.

increase when coupled with the prospect of shaming. This logic is outlined in the *Conditional International Norms and Shaming Hypothesis*.

> **Conditional International Norms and Shaming Hypothesis**: The positive effect of reminding state agents about international norms is greater when the requester has a higher ability to shame the state than when they have a lower ability.

3 Research Design

We tested our theoretical expectations by conducting an audit experiment (or online field experiment) with a sample of foreign missions. The benefit of our research design is that it allows us to identify the causal effect of norms and the possibility of shaming, while also examining how state agents respond to these factors.

3.1 Sample

Our population of interest consists of the 1,098 foreign missions and consulates general in Canada, the United States, and the United Kingdom.[4] We selected missions and consulates in these countries because they exhibit similarly high levels of state respect for human rights (Fariss, 2014). The idea here is to observe how bureaucrats from foreign regimes respond to international norms and the threat of sanctioning in contexts where the public and political elites exhibit strong support human rights issues. In such places, we would expect the power of norms and potential sanctions to be at their highest. In that sense, we conduct an 'easy test' of our theory: if the relationships that we are looking for exist, we should find them in these three countries.

We attempted to collect email addresses for this census of missions. Many missions, however, do not provide general email addresses for public correspondence. We exclude these from our sample. Some missions, on the other hand, list the same email address as

[4]There are 128 countries represented by 277 foreign diplomatic missions in Canada. There are 176 countries represented by 693 foreign diplomatic missions in the United States. There are 164 countries represented by 191 foreign diplomatic missions in the United Kingdom.

other missions. In this case, we randomly dropped all but one of the missions from our sample. Additionally, some missions either listed obviously incorrect email addresses (e.g., they do not contain domain extensions, or text before the symbol) or provide email addresses that did not work. We also excluded those from our sample. Our final sample then consisted of 669 missions. It consists of 246 embassies and 423 consulates that cover 155 countries in the 3 host countries in our sample (i.e. Canada, United Kingdom, and United States).

Before sending emails to these missions, we spoke with staffers at several of them. Our two primary goals in these conversations were to (1) determine who replies to emails like the ones we were about to send and (2) assess the believability of our email requests. We learned that emails are generally answered by foreign staff, particularly in consulates, where local national staff numbers are often small. We also learned that foreign missions receive many emails every day dealing with a wide variety of subjects and each typically receives a reply. Based on the descriptions of these emails, we have little reason to think that ours stood out as odd. This is even more likely because our understanding of the daily workflow for most staff suggests that had little time to assess the believability of our emails.

3.2 Experimental Design

The study occurred over three waves. In the first wave, we emailed foreign missions with a basic request. We did this to establish a baseline response rate for the units in our sample, which we could then use to place the experimental response rate in context. Figure 1 presents the email text used in wave one. While we kept the text of the email the same, we randomized the separators (i.e. white space), valediction, and sender identity. We did this to preserve participant naivety and to minimize the possibility of spillover. We delivered these emails from April 27-28, 2016.

In the second wave, we sent each mission a request for information about contacting domestic prisoners. We selected this topic for our correspondence because according to the United Nations Standard Minimum Rules for the Treatment of Prisoners, prisoners have the

Figure 1: Wave One Email Text

```
Hello,
Could you send me information about how to obtain a copy of your
constitution?
<separator>
<valediction>,
<name>
<separator>
```

Note: Text in <> represents randomly assigned elements.

right for individuals to contact them. Each email contained this request along with a set of randomized cues designed to test our theoretical expectations.

To test our four hypotheses, we used a $2x2$ factorial design. The first factor varies whether or not we include a norms cue. This is indicated by language that reminds the recipient of the acknowledged right that prisoners have to communicate with the outside world. The idea here is that reminding recipients of this norm might increase norm compliance.[5]

The second factor varies the extent to which we signal our ability to shame the recipient for disregarding our request. This is indicated by whether the putative sender works as a social sciences teacher or for a fictional human rights advocacy group. The idea here is that an advocacy group has a greater ability to highlight, and possibly shame, a recipient's unhelpfulness than a sender in some other line of work, such as a teacher.[6]

Figure 1 presents the cues. As in the first wave, we randomized the separators, valediction, and sender identities.[7] Figure 2 presents the email text used in wave two. We sent these emails on May 2, 2016.

[5]One possible concern might be that staff misinterpret our language and assume that this a legal requirement instead of an international norm. If that is the case, though, we would expect to see higher levels of compliance for missions that are assigned this treatment than we report below.

[6]One concern might be that we do not specify the name of the HRO in our email. We did not do this because we wanted to be able to separate the effect of shaming capacity from any particular effects related to specific HROs. For example, if we would have included language suggesting that we were a research officer with Amnesty International, we would not have been able to determine if any observed effect was because Amnesty International has the capacity to shame regimes or for some other idiosyncratic reason related to Amnesty International. This design choice, though, potentially means that the generalizability of our results regarding this cue are limited.

[7]The randomization of sender identity was constrained so that we did not use the same identity in both wave one and wave two emails.

Table 1: Request texts

Subject line	Question
Shaming Cue	a research officer with a human rights advocacy group,
No Shaming Cue	social studies teacher who teaches about human rights advocacy,
Norms Cue	Rule 37 of the United Nation's' Standard Minimum Rules for the Treatment of Prisoners' suggests that this is possible.

Figure 2: Wave Two Email Text

```
To Whom it May Concern,
My name is <name>.  I'm a <shaming cue> and I have a long-standing interest
in prisoner well-being.  I would like to write to some prisoners in [country
name].  <norms cue> How can I do this?
<separator>
<valediction>,
<name>
<separator>
```

Note: Text in <> represents randomly assigned elements.

We block randomized our treatments based on a key pre-treatment covariate plausibly predictive of mission responsiveness — mission location (i.e. the mission's host country) (Moore and Schnakenberg, 2012). The intuition here is that missions might respond at different rates across these countries because of local norms, such as expectations about political elite communication. One might also want to block on mission type (i.e. embassy or consulate) but there is not enough variation in this attribute within host countries.

In the third wave, which occurred a week after the first and second waves, we emailed all missions with a reminder about our request to contact prisoners. We sent them the same emails as in wave two but added to the top of each email a sentence stating that we were following up on our previous request.[8] We did this because our pre-experiment interviews with diplomatic staff suggested that since they receive many requests we should be persistent

[8]One might think about sending the follow up emails only to missions who did not yet reply. The problem with this, however, is that it would make the reminder conditional on treatment assignment. We wanted to avoid this unnecessary complication. An objection to our approach, however, is that the reminder email might seem odd to missions who *did* reply. We think, however, that mission personnel would have been understanding of this given that individuals often overlook emails and that many email replies are erroneously sent to SPAM folders.

with our requests. Figure 3 presents the email text used in wave three. We sent these emails on May 6, 2016.

Figure 3: Wave Three Email Text

```
To Whom it May Concern,
Since I haven't received a response to my inquiry, I wanted to send a
follow up email.  My name is <name>.  I'm a <shaming cue> and I have a
long-standing interest in prisoner well-being.  I would like to write to
some prisoners in [country name].  <norms reminder cue> How can I do this?
<separator>
<valediction>,
<name>
<separator>
```

Note: Text in <> represents randomly assigned elements.

After sending the third wave emails, we waited two weeks to receive replies. We then collected our outcome measures. As indicated in our pre-registration plan, we collected two outcome measures. The first, Email Response is coded 1 if missions reply to our email and 0 otherwise. The second, Helpful Email, is coded 1 if missions send a helpful reply to our email and 0 otherwise. We define a helpful reply as one that provides actionable information. A common example of this is contact information for a national Prison and Probation Service headquarters. A typical example of an unhelpful reply is one that advises us to consult Google.

Prior to analyzing the results, we checked the integrity of the randomization procedure by examining whether mission location is balanced across treatment groups. We tested this by comparing the log-likelihood statistics of a null model and a full model that regressed treatment assignment on mission location (Gerber and Green, 2012, 107). The results suggested that we cannot reject the null hypothesis that the full model fits the data better ($p \approx 1$). This means that covariate imbalance is no larger than might occur by chance (Gerber and Green, 2012), which is an indicator of successful randomization. We now turn to the presentation of the empirical results.

4 Results

The overall response rate for our experimental emails was 29.30 percent. This is slightly higher than the response rate for our baseline emails (24.36 percent). In general, this reply rate is in line with response rates from audit experiments with elites (Costa, N.d.). Taken together, these two facts suggest that our emails were taken at face value.

As expected, the response rates differ by treatment condition. Table 2 presents the raw response rates. Shaming denotes emails that received the shaming condition, Norms denotes emails that received the norms reminder condition, and Norms and Shaming denotes emails that received both conditions. The table shows that compared to the control email, responses were slightly higher for emails that contained the shaming cue and much higher for emails that contained the norms cue. Indeed, the raw response rate for emails with the norms cue is 186 percent the raw response rate for control emails. These two patterns are in line with our theoretical expectations. On the other hand, the response rate is slightly *lower* for emails with both the norms and shaming cues, which runs counter to our expectations.

Table 2: Raw Response Rates

Control	Shaming	Norms	Norms and Shaming
0.247	0.251	0.460	0.216

Table 2 presents the raw response rates across our experimental conditions. Shaming denotes emails that received the shaming condition, Norms denotes emails that received the norms reminder condition, and Norms and Shaming denotes emails that received both conditions.

To evaluate whether these differences in response rates are statistically significant, we first conduct an omnibus Wald test of the null hypothesis that none of the treatments has any effect. This tells us that we can reject the null that email replies are independent of our 4 experimental treatments ($p \approx 0$). This means that our treatments have a statistically significant effect on the probability of receiving a reply from missions.

To investigate the effect of individual treatments, we estimate a linear probability model

(LPM). We use an LPM because estimates are unbiased if the model is correctly specified and because the results are easy to interpret (Wooldridge, 2010).[9] We can obtain the correct specification by including only treatment indicators, the blocking covariate, and fixed effects for the other randomized aspects of our email texts.[10] The reference category for this model is the control condition that contains no reminder about international norms and comes from a putative social studies teacher. To account for heterogeneity in the error term, we use Bell-McCaffrey adjusted standard errors as recommended by Lin and Green (2015).[11]

Figure 4 displays the results of this model.[12] Black plotted points represent estimated coefficients, thick red bars represent 90 percent confidence intervals, and thin red bars represent 95 confidence intervals. The gray dotted line denotes 0. The size and direction of the estimated coefficients are in line with the raw response patterns in Table 2. The estimates for the shaming treatment is positive, as we would expect, but is estimated imprecisely. We cannot reject the null that our shaming cue has no effect on state agent compliance.[13]

On the other hand, there appears to be strong evidence for the *International Norms Hypothesis*. This is indicated by the positive and statistically significant coefficient on the NORMS coefficient ($p < 0.01$). The effect of this cue is substantively meaningful. It increases the probability of receiving a reply by approximately 21 percent. This means that a basic reminder about the existence of an internal norm has a large, important effect on state compliance.

[9] Since our outcome measure is bound between $[0 - 1]$, we could also use logit or probit models to estimate effects. The results from these models are substantively similar to estimates from the LPM model.

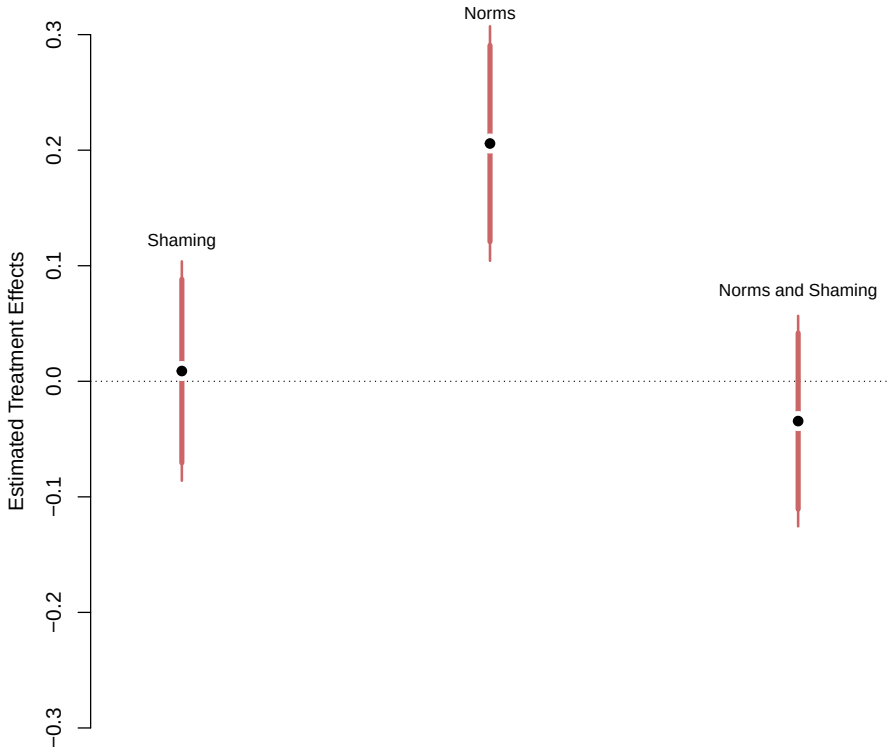
[10] As described above, these include the wave in which the email was sent and the valediction and separator used in the email. The inclusion of covariates or fixed effects has very little effect on point estimates. This is typically true of randomized experiments with large sample sizes (Gerber and Green, 2012).

[11] Our results are substantively the same if we use classic standard errors or if we use HC2 robust standard errors (Angrist and Pischke, 2008).

[12] Appendix A presents the LPM coefficient estimates and BM standard errors.

[13] There are potentially many reasons for this. One might be that mission staff did not believe our emails actually came from an employee at an HRO. If this was the case, then they would have not worried about our potential power to sanction them for a non-reply. If this were true, though, we should see that the response rate for emails with the shaming cue would be much lower than the response rate for emails from the putative social studies teacher. The idea here is that personnel should be less likely to reply to emails of questionable providence than realistic emails. We do not see this pattern, however. In fact, though, there is essentially no difference in these reply rates.

Figure 4: Treatment Effects on Email Response



Note: This figure displays the results of an LPM with Email Response as the outcome measure. The model includes treatment indicators (i.e. Shaming, Norms, and Norms and Shaming), the blocking covariate, and fixed effects for the other randomized aspects of our email texts. Black plotted points represent estimated coefficients, thick red bars represent 90 percent confidence intervals, and thin red bars represent 95 confidence intervals. The gray dotted line denotes 0.

Contrary to our expectations, though, the joint effect of these cues is lower than the effect of the NORMS cue. In other words, the norms cue on its own does more to increase compliance than the norms cue combined with the sanctions cue. One possible explanation for this counter-intuitive finding is that when the treatments are combined it creates a 'backfire' effect. The intuition here is that when missions receive both cues, they might prefer not to respond at all than to reply in a way that will not satisfy the research officer of a human rights advocacy group.

Since these results depart from our *a priori* expectations, we investigate whether they are a statistical fluke by conducting a Studentized permutation test (Gerber and Green, 2012). The p-value from this test is the fraction of test statistics strictly larger than the test statistic in our sample (Lin and Green, 2015). The results from these tests provide additional confidence to our findings reported above.[14]

Next we examine whether our treatments effect our other outcome measure — HELPFUL RESPONSE. As a reminder, this outcome is coded 1 if we receive a helpful response and 0 otherwise, and we defined a helpful reply as one that provides actionable information. In effect, our coding rule means that we recode EMAIL RESPONSE so that unhelpful replies are coded as 0. Approximately 20 percent of the missions sent us a helpful response.

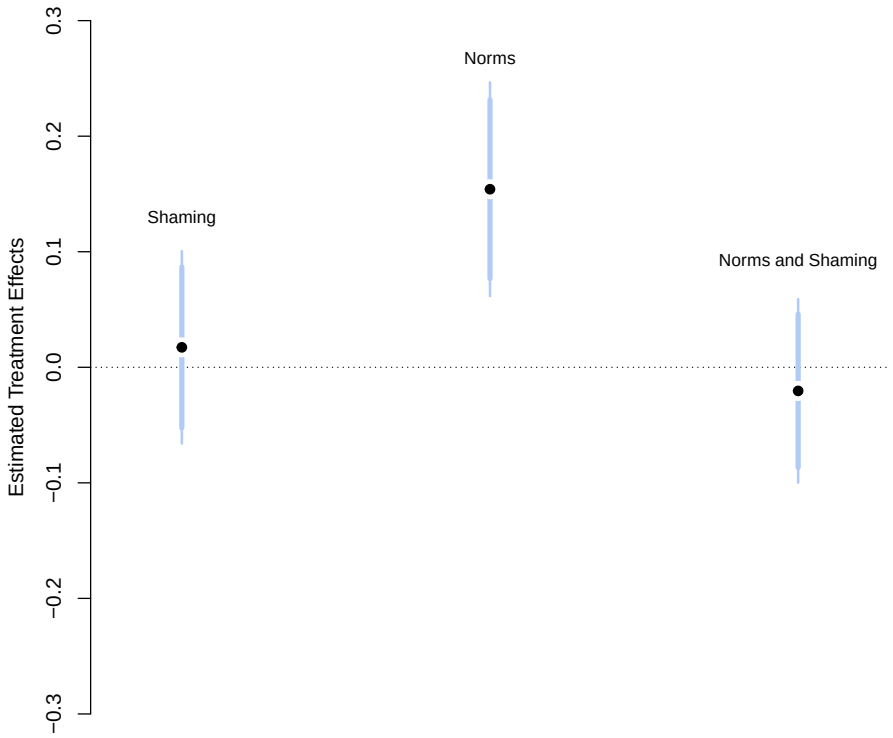
The raw helpful reply rates mirror the email response rates. Emails that contained the shaming or norms reminder cues were more likely to receive a helpful reply, while emails that contained the compound treatment were less likely to receive a helpful reply. An omnibus χ^2 test suggests that the helpful reply rates are not independent of our treatments ($p \approx 1$).

We again investigate the effect of individual treatments by estimating an LPM with BM-adjusted standard errors. Figure 5 displays the results of this model.[15] Black plotted points represent estimated coefficients, thick blue bars represent 90 percent confidence intervals, and thin blue bars represent 95 confidence intervals. The gray dotted line denotes 0.

[14] Since we have a large sample and use a (modified version of) robust standard errors, this is not surprising (Gerber and Green, 2012).

[15] Appendix B presents the LPM coefficient estimates and BM standard errors.

Figure 5: Treatment Effects on Helpful Response



Note: This figure displays the results of an LPM with Helpful Response as the outcome measure. The model includes treatment indicators (i.e. Shaming, Norms, and Norms and Shaming), the blocking covariate, and fixed effects for the other randomized aspects of our email texts. Black plotted points represent estimated coefficients, thick blue bars represent 90 percent confidence intervals, and thin blue bars represent 95 confidence intervals. The gray dotted line denotes 0.

Again we find little evidence for our *Shaming Hypothesis*; nor is the *Conditional International Norms and Shaming Hypothesis* supported. This is indicated by the imprecise estimates on the SHAMING and NORMS AND SHAMING cues. There is, however, strong evidence for the *International Norms Hypothesis*. This is indicated by the large and statistically significant NORMS coefficient ($p < 0.01$). As before, we implement a Studentized permutation test, which provides additional support for these results.

Treatment Effect Heterogeneity

We expected treatment effects to be heterogeneous. In particular, we anticipated that the treatment effect would be lower for missions from autocratic countries than for missions from democratic countries, as well as for missions from states that routinely violate human rights than states that do not. We find little evidence, however, that these country-level characteristics influence the effect of treatments on either of our outcome measures. As pre-specified, we also investigate in an exploratory fashion whether treatment effects vary as a function of the global region in which states are located. Again, we find little evidence that treatment effects are conditional on these characteristics. We interpret the lack of evidence for treatment effect heterogeneity as a sign that the strategic calculus of mission responsiveness is widely shared.

5 CONCLUSION

Having conducted an online field experiment with foreign missions, we find strong evidence for the effect of international norms on state respect for prisoners' rights. A simple reminder about the norm significantly increased both mission response rates as well as the helpfulness of the response. However, we find little evidence for an independent effect of shaming.

Most surprisingly, and counter to our expectations, we find that when we combine the norms and shaming cues, the effect is much smaller (and in opposite sign) than the effect for norms alone. This suggests a 'backfire' effect, where mission personnel who received the

shaming and norms cues were deterred from replying. We speculate that when bureaucrats at foreign missions receive both cues, they might prefer not to respond than to reply in a way that might be unsatisfying to an individual with potential shaming power. In effect, a non-reply is less risky to the career of a bureaucrat than a reply.

The results from our experiment contribute to the human rights literature by providing empirical evidence for the effect of international norms on the behavior of state agents. This evidence is consistent with the constitutive model of international law and suggests that international institutions can affect state behavior by creating self-enforcing norms. Furthermore, these findings provide partial evidence for the effect of international norms and the possibility of shaming on the individual level. This suggests that researchers should devote more attention to street-level bureaucrats, who are commonly tasked with implementing state policy. For this reason, and because state agents are typically the individuals that citizens are most likely to interact with, they play a central role in whether human rights are respected in practice. Finally, the results suggest that state compliance with some human rights norms is not always conditional on such state characteristics like regime type or human rights record.

These findings have clear policy implications for HROs and states seeking to enforce human rights norms. In particular, our finding on the joint effect on norms and shaming highlights an important proviso for human rights advocates that there may be unintended and deliterious consequences to efforts aimed at inducing compliance if they increase the risk to individual agents. This finding can possibly explain why some HROs prefer to use soft tactics over hard pressure on states and their agents in order to change their behavior. While our explanation of this finding remains speculative, it suggests that future research should investigate the parameters under which state agents are willing and able to respect rights. In addition, the empirical evidence for the effect of international norms suggests the importance of addressing human rights issues through international law and international community pressure in improving human rights practices across states.

References

Abbott, Kenneth W, Robert O Keohane, Andrew Moravcsik, Anne-Marie Slaughter and Duncan Snidal. 2000. "The concept of legalization." *International organization* 54(03):401–419.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Ausderan, Jacob. 2014. "How naming and shaming affects human rights perceptions in the shamed country." *Journal of Peace Research* 51(1):81–95.

Benhabib, Seyla. 2009. "Claiming rights across borders: International human rights and democratic sovereignty." *American Political Science Review* 103(04):691–704.

Brysk, Allison. 1993. "From above and below social movements, the international system, and human rights in Argentina." *Comparative Political Studies* 26(3):259–285.

Buchanan, Allen E. 2007. *Justice, legitimacy, and self-determination: moral foundations for international law*. Oxford University Press on Demand.

Bush, Sarah Sunn. 2011. "International politics and the spread of quotas for women in legislatures." *International Organization* 65(01):103–137.

Costa, Mia. N.d. "How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials." . Forthcoming.

Creamer, Cosette D. and Beth A. Simmons. 2015. "Ratification, Reporting, and Rights: Quality of Participation in the Convention against Torture." *Human Rights Quarterly* 37(3):579–608.

Dancy, Geoff and Christopher J Fariss. 2017. "Rescuing Human Rights Law from International Legalism and Its Critics." *Human Rights Quaterly* 39:1–36.

Davis, David R, Amanda Murdie and Coty Garnett Steinmetz. 2012. "" Makers and Shapers": Human Rights INGOs and Public Opinion." *Human Rights Quarterly* 34(1):199–224.

Dietrich, Simone and Amanda Murdie. 2015. "Human rights shaming through INGOs and foreign aid delivery." *The Review of International Organizations* pp. 1–26.

Fariss, Christopher. 2014. "The Changing Standard of Accountability and the Positive Relationship between Human Rights Treaty Ratification and Compliance." *British Journal of Political Science* .

Finnemore, Martha and Kathryn Sikkink. 1998. "International norm dynamics and political change." *International Organization* 52(4):887–917.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Goldsmith, Jack and Stephen D Krasner. 2003. "The limits of idealism." *Daedalus* 132(1):47–63.

Hafner-Burton, Emilie M. 2008. "Sticks and stones: Naming and shaming the human rights enforcement problem." *International Organization* 62(4):689–716.

Hill, Daniel W. 2010. "Estimating the effects of human rights treaties on state behavior." *The Journal of Politics* 72(4):1161–1174.

Keck, Margaret and Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Ithaca, NY: Cornell University Press.

Krain, Matthew. 2012. "Jaccuse! Does Naming and Shaming Perpetrators Reduce the Severity of Genocides or Politicides?" *International Studies Quarterly* 56(3):574–589.

Lebovic, James H. and Erik Voeten. 2009. "The Cost of Shame: International Organizations and Foreign Aid in the Punishing of Human Rights Violators." *Journal of Peace Research* 46(1):79–97.

Lewis, Paul G and S Karthick Ramakrishnan. 2007. "Police Practices in Immigrant-Destination Cities Political Control or Bureaucratic Professionalism?" *Urban Affairs Review* 42(6):874–900.

Lin, Winston and Donald P Green. 2015. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." *Berkeley. Retrieved from www. stat. berkeley. edu/˜ winston/sop-safety-net. pdf (2014). Promoting transparency in social science research. Science (New York, NY)* 343(6166):30–1.

Lipsky, Michael. 1980. "Street-Level Bureaucracy." *New York* .

Lipsky, Michael. 2010. *Street-level bureaucracy, 30th ann. Ed.: dilemmas of the individual in public service*. Russell Sage Foundation.

Marrow, Helen B. 2009. "Immigrant bureaucratic incorporation: The dual roles of professional missions and government policies." *American Sociological Review* 74(5):756–776.

Meernik, James, Rosa Aloisi, Marsha Sowell and Angela Nichols. 2012. "The impact of human rights organizations on naming and shaming campaigns." *Journal of Conflict Resolution* 56(2):233–256.

Meier, Kenneth J and Laurence J O'Toole. 2006. "Political control versus bureaucratic values: Reframing the debate." *Public administration review* 66(2):177–192.

Moore, Ryan T and Keith Schnakenberg. 2012. "blockTools: Blocking, assignment, and diagnosing interference in randomized experiments." *Version 0.5-6, August* .

Morgan, Stephen L and Christopher Winship. 2014. *Counterfactuals and causal inference*. Cambridge University Press.

Murdie, Amanda and Dursun Peksen. 2013. "The impact of human rights INGO activities on economic sanctions." *The Review of International Organizations* 8(1):33–53.

Murdie, Amanda M. and David R. Davis. 2012. "Shaming and blaming: Using events data to assess the impact of human rights INGOs." *International Studies Quarterly* 56(1):1–16.

Pearl, Judea. 2009. *Causality*. Cambridge University Press.

Terman, Rochelle and Erik Voeten. 2017. "The relational politics of shame: Evidence from the universal periodic review." *The Review of International Organizations* pp. 1–23.

Wexler, Lesley. 2003. "The International Deployment of Shame, Second-Best Responses, and Norm Entrepreneurship: The Campaign to Ban Landmines and the Landmine Ban Treaty.".

Wilson, James Q. 1989. "Bureaucracy: What government agencies do and why they do it.".

Winship, Christopher and Stephen L Morgan. 1999. "The estimation of causal effects from observational data." *Annual Review of Sociology* 25(1):659–706.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Supplementary Information for

"Evaluating the Influence of International Norms and Shaming on State Respect for Rights: A Field Experiment with Foreign Embassies"

(Not for publication.)

Appendix A

Table 3

	Coefficient Estimate	Standard Error	t-stat
Shaming	0.009	(0.048)	0.183
Norms	0.206	(0.052)	3.974
Norms and Shaming	-0.034	(0.046)	-0.741

Note: This table displays the results of an LPM with Email Response as the outcome measure. The model includes treatment indicators (i.e. Shaming, Norms, and Norms and Shaming), the blocking covariate, and fixed effects for the other randomized aspects of our email texts. The left column contains estimated coefficients from this model, the center column contains standard errors in parentheses, and the right column contains t-stats.

Appendix B

Table 4

	Coefficient Estimate	Standard Error	t-stat
Shaming	0.017	(0.042)	0.406
Norms	0.154	(0.047)	3.265
Norms and Shaming	-0.020	(0.041)	-0.505

Note: This table displays the results of an LPM with Helpful Response as the outcome measure. The model includes treatment indicators (i.e. Shaming, Norms, and Norms and Shaming), the blocking covariate, and fixed effects for the other randomized aspects of our email texts. The left column contains estimated coefficients from this model, the center column contains standard errors in parentheses, and the right column contains t-stats.