Note to Readers: Thank you for taking the time to read this early-stage proposal. The goal of this research is to identify if and how strategies of “naming and shaming” mobilize policy change. These strategies are often used by NGOs and are a key component of norm socialization theories in IR. Nevertheless, it is unclear if and how they influence elites. I made contact with the Ministry of Foreign affairs in Israel in order to conduct a survey experiment on trainees of their cadets course. By the end of May 2019, I intend to conduct a pilot experiment on the Israeli public in order to test my hypotheses and make adjustments. Hopefully, during June-July 2019 the elite experiment will take place. I would be grateful for any feedback, especially on the emotion-eliciting treatments.
“Human Rights change starts with the facts... We use our analysis to influence and press decision makers to do the right thing.”

– Amnesty International

1 Introduction

Do publications of norm violations cause policy change? Publicizing states’ violations of international norms in order to promote a change in policy is a popular strategy, particularly endorsed by international and non-governmental organizations. In 2018 alone Amnesty International, a non-governmental organization focused on human rights, has published over 950 reports about violations conducted by various states\(^1\), with the objective of influencing governments to change their policies. Nevertheless, the literature provides conflicting evidence when measuring the causal relationship between naming norm violators and policy change. Moreover, little theoretical work has been done in order to account for the mechanisms through which change is mobilized. Who is being targeted? What motivates targets to change their course of action? How?

Drawing on the literature on shame and guilt in Psychology, this study offers a new theoretical framework to conjecture how leaders may react to strategies of ‘naming and shaming’ of their nation. I argue that leaders who experience guilt will be more likely to change policy to conform with the international norm, while those who experience shame will be more likely to defend current policies. These feelings are moderated by levels of political self-efficacy. In order to account for this relationship and overcome issues of identification and measurement, I propose to conduct an elite survey experiment in Israel. I highlight the need of a new theoretical framework as well as a novel empirical strategy that accounts for the causal relationship between publicity and policy change. I begin by reviewing the current literature and presenting the framework I propose. I then describe the research design and its limitations.

2 Contribution

2.1 Micro foundations of norm-socialization

Norm socialization is arguably the most prominent process in Constructivist thought and perhaps one of the paradigm’s greatest contributions to International Relations theory (Adler 2013). ‘Naming and shaming’ – the process of publicizing states’ violations of international norms, is viewed as part of this socialization process (Risse-Kappen et al. 1999). Norm entrepreneurs such as NGOs, IOs, states or other individual actors use this tactic in

\(^1\)Amnesty International Research
order to pressure repressive governments to change their policy. Naming therefore promotes international norms by reinforcing shared understandings in the international community (Krains 2012). States are ultimately socialized into internalizing certain moral standards and changing their domestic policies accordingly (Finnemore and Sikkink 1998).

While this theory provides a compelling argument at the international level of analysis, it provides little theory to explain how change is mobilized at the domestic level. In fact, it disregards altogether the role of decision makers within this process and simply assumes that officials are pressured into complying. Constructivism does not unravel the mechanisms at the individual level of analysis or provides a clear theory to explain why (and when) leaders may change policy following publication. In recent years, IR literature has taken a behavioral turn with a focus on micro foundations and political psychology (Kertzer 2017). A variety of studies unravel the individual mechanisms that could shed light on larger international processes. In this research, I propose a theoretical framework that deals with the micro foundations of international norm socialization. I intend to unravel the ways in which decision makers play a role in implementing international norms, when their state’s violations are put in the global spotlight. Since shame is a human emotion, it seems rather intuitive to consider the ways in which it is internalized at the individual level, rather than simply assuming that it shall effect state behavior as a whole.

2.2 Need of experimental evidence

Recent IR literature offers conflicting results when measuring the effects of naming and shaming. A number of studies have found a strong relationship between public naming and policy change. Using data on criticism of political repression in Latin America, Franklin (2008) finds an impact of public shaming on policy change. Murdie and Davis (2012b) find that states improve human rights practices after they have been publicly shamed by human rights organizations (HROs) physically present in targeted states. Kelley and Simmons (2015) show that numerical indicators can be used by international actors as shame-inducing tools, and argue that governments are more likely to criminalize human trafficking when they have been placed on a “watch list” or included in the US annual trafficking persons report.

However, other studies find little to no effect of public shame on policy change. Bai-Finnemore and Sikkink (1998) argue in their seminal work that states conform with international norms in the second stage of the “norm life cycle” due to concerns of legitimacy. This theory does not deal with decision makers (but rather reflects on the state as a whole), nor does it mention reputation or legitimacy as concerns that arise from “naming and shaming”. However, it could be applied as a potential mechanism. Hence, in section 3.2.2 I present this logic as an alternative explanation.
ley (2008) shows that shaming tactics did not help socialize states into ending commercial whaling, but rather provoked a domestic backlash. Using data on governments who have been publicly named by the United Nations Commission on Human Rights (UNCHR), Lebovic and Voeten (2009) provide some empirical support for the existence of mechanisms through which governments are held accountable for their human rights violation. However, they are unable to show a clear change in policy. Hafner-Burton (2008) shows that governments put in the global spotlight rarely stop/lessen acts of immorality, but rather often increase them after publicity. Finally, Carnegie and Carson (2018) recently argued that shaming causes instability and undermines the normative legal order.

There is reason to speculate issues with identification and measurement have led to these contradictions in the literature. First, there are multiple variables that predict both shaming and outcomes of interests. Since publicity is often a result of new information of violations, countries that seek to improve their human rights conditions tend to be more transparent with respect to their violations, thus making it difficult to disentangle the causal direction of the effect. Second, measurement strategies of policy change vary between studies and often suffer from issues of content validity as well as missing values (Fariss 2018). These issues call for an experimental design that will allow us to account for this causal relationship.

3 Theoretical Framework

Why, if so, would policy makers conform with international norms following publication? Two logics are applied to explain this relationship: Emotional and Rational. Each explanation provides a different mechanism to explain the relationship between publicity and policy change. The rational explanation could be viewed as more conventional, with perhaps stronger roots in existing IR theories. It does not, however, thoroughly explain why and when publication will not lead to change. The emotional explanation is the main theoretical contribution of this paper, borrowed from theories of social emotions in Psychology. I present both of these logics and derive testable hypotheses.

3.1 Emotional

The term used to describe the strategy of publicizing states’ violations – ‘naming and shaming’, suggests that emotions play a role within this process. States, or state officials, are supposedly ashamed and therefore change their actions. Psychology, however, describes shame as a defensive, rather than reparative, emotion. It is therefore unlikely that individuals that experience shame will improve their actions as a result. In this sub-section I conjecture how emotions may play a role in determining the effects of naming, by drawing on the differences between Shame and Guilt.
Shame and Guilt are both defined as emotions that arise when actors act in ways that violate moral standards and values. These emotions take place when there is discrepancy between how one is expected to behave and how one actually behaves (Doosje et al. 1998; Tangney, Stuewig and Mashek 2007). Scholars most commonly differentiate guilt and shame based on the degree to which the person interprets the emotion-eliciting event as a failure of self or behavior (Tangney, Stuewig and Mashek 2007). According to Lewis (1971) the emphasis in shame is on the self (I did that horrible thing), while the emphasis in guilt is on the behavior itself (I did that horrible thing). Guilt is therefore concerned with a particular behavior and not with one’s core identity. Shame, on the other hand, brings one’s self-image and self-esteem into question (Lindsay-Hartz 1984). Hence, shame is considered a much more painful emotion, found to promote defensiveness and interpersonal separation. Guilt, on the other hand, appears to motivate reparative action, foster empathy for the “other” (as it is not the self who is evaluated) and promote constructive change.

These emotions, however, are concerned with the actions that individuals conduct themselves. Since “naming and shaming” is concerned with a state or a government, it could be argued that these emotions cannot be applied to individuals, as they would be experiencing them on behalf of their nation. However, studies in Psychology show that individuals experience these emotions on a collective level as well.

Social Identity Theory (SIT) suggests that people derive a sense of identity from social associations and are motivated to maintain a positive image of their social groups (Tajfel 2010). Hence, when faced with transgressions of their in-group, individuals may experience collective guilt or collective shame. SIT suggests that individuals will feel collective guilt when they believe that they themselves had some control over the blameworthy event, or if they believe they can control its continuation or implications. Collective shame, on the other hand, is often experienced when actors feel like they have no control over the blameworthy event but fear that these actions reflect poorly on the image of their in-group, and therefore – themselves. Similar to individual shame and guilt, collective shame encourages defensive action while collective guilt encourages reparative action (Lickel, Schmader and Barquissau 2004).

If so, in collective guilt, individuals believe that the wrongs of their group implicate something about their own behavior, and are therefore motivated to change it. In collective shame, people believe that the wrongs of their group say something about the very nature of who they are, and are therefore more likely to defend their in-group and their own

---

3 Arguably, decision makers should be more accountable for these actions and may therefore experience them at the individual level. Nevertheless, the decision making process is usually a collective one, and there is therefore a need to discuss a theory that deals with these emotions at a group level.
self-esteem. These feelings are moderated by actors’ perception of their control of the blameworthy event. The differences between these emotions are summarized in Table 1. A range of empirical studies have found evidence to support these conjectures, often dealing with ethnic, gender and other social groups within the state (Brown et al. 2008; Brown and Cehajic 2008; Gunn and Wilson 2011). I propose applying this logic to international politics in order to understand the ways in which naming and shaming may encourage policy change. The following hypotheses are derived from this theory.

**H1a:** Publication of a state’s violation is more likely to trigger collective guilt in leaders with high political self-efficacy. When a nation is named, its transgressions are sharpened and brought forth. Leaders faced with these accusations are likely to feel guilty if they believe that they have some control over the blameworthy event or its continuation. Such belief is often referred to in literature as political self-efficacy (Caprara et al. 2009a) – the degree to which individuals believe in their own capabilities to control political change. If leaders strongly believe that they have the power to control policy, they are likely to believe that the nation’s wrongdoings reflect their own behavior. This is then followed by **H1b:** Collective guilt will encourage leaders to conform with moral standards. Since leaders are concerned with their in-group’s behavior, they are more likely to change it accordingly.

**H2a:** Publication of a state’s violation is more likely to trigger collective shame in leaders with low political self-efficacy. These leaders believe that they have no control over their in-group’s actions and therefore perceive publication as a threat to the very essence of their group. For them, flawed aspects of their social identity compromises their own self-esteem and self-image (as it is partially derived from their social association). Hence, it is more likely that these leaders will experience collective shame. According to SIT, collective shame is accompanied by defensiveness, which implies that – **H2b:** Collective shame will encourage leaders to defend the states’ actions and to refuse to conform with moral standards. As collective shame triggers defensiveness, I conjecture that leaders will defend their nation’s actions and therefore refrain from changing policy to conform with moral standards.

To summarize, when violations of norms are publicized (X), leaders’ reactions alter based on whether they feel guilty or ashamed. Leaders who experience guilt (G) will be

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Eliciting Event</th>
<th>Subject of Focus</th>
<th>Actor’s Perception</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guilt</td>
<td>In-group transgression</td>
<td>Behavior</td>
<td>Control over event</td>
<td>Reparative action</td>
</tr>
<tr>
<td>Shame</td>
<td>In-group transgression</td>
<td>Self and Self-esteem</td>
<td>No control over event</td>
<td>Defensive action</td>
</tr>
</tbody>
</table>

Table 1: Guilt and Shame
more likely to change policy to conform with the international norm \((Y_1)\), while those who experience shame \((S)\) will be more likely to defend current policies \((Y_2)\). Feelings of guilt and shame are moderated by the degree to which actors believe in their own capabilities to control political change \((M)\). This causal process is outlined in Figure 1.

### 3.2 Rational

Two logics could be applied as alternative explanations to the emotional one. I define these as rational explanations as they are more clearly based on reason and logic and could be seen as goal-oriented. The first is rather intuitive and simple, the second serves as a stronger alternative explanation and is more deeply rooted in current IR theories.

#### 3.2.1 Information

One alternative explanation to the theory previously outlined is that leaders are not driven by emotions but rather by revelation of new information. Assuming that actors did not know of their nation’s violation, it is possible that they decided to change policy in light of new information. While this theory is plausible, it seems rather unlikely that decision makers are unaware of their country’s wrongdoings. Nevertheless, it is crucial to test this potential explanation.

#### 3.2.2 Reputation

The second rational explanation is not concerned with information but rather with actors’ concern with their state’s reputation. In IR, reputation is often understood as “a belief about a trait or behavioral tendency of an actor, based on that actor’s past behavior” (Dafoe, Renshon and Huth 2014). An actor could have a bad reputation or a good one, commonly known as “prestige”. Reputation is not absolute but rather relative to peers (David-Barrett and Okamura 2016), it is therefore often considered to foster competition amongst actors and a desire to do better (Kelley and Simmons 2015).

In the norm cascade theory, it is often suggested that actors first comply with norms because of reputational and competitive concerns. A sincere internalization only follows
in later stages of the norm’s life cycle, and is often seen as a result of habit and practice (Finnemore and Sikkink 1998). Taking this into account, it could be argued that governments should respond to “naming and shaming” in order to protect their reputation and legitimacy. If so, according to this logic, publications about failure raise reputational concerns amongst leaders, causing them to change their policy in an attempt to do better than (or just as well as) other actors.

While this theory provides a compelling explanation as well, it provides less information about cases in which publication may not work. In other words, why would some countries adhere to normative behavior following publications while others do not? Why does reputation foster change in some cases, but not in others? If actors are always concerned with reputation, one may argue that following a revelation of violations all (or at least most) violations will be targeted. The emotional framework provides explanations for both cases in which emotions lead to compliance (guilt) or violation (shame). It therefore does a better job in discerning the different behaviors of actors. Nevertheless, the reputation explanation is an important one that cannot be disregarded.

4 Research Design

4.1 Treatments

The main independent variable that will be explored in this research is naming norm violators. In order to account for this variable, elites will be randomly assigned a text which includes some kind of critical report of their country’s transgressions. To account for mechanisms and minimize bundled effects, I suggest assigning elites with three different texts that embody competing explanations for the hypothesized relationship. Control groups will be assigned an unrelated text.

4.1.1 Normative Information (X)

Israeli elites will be assigned an Amnesty International report which condemns Israel’s policies toward Palestinian prisoners. This is a text similar to others, published by a non governmental organization in order to report Israel’s immoral activity (“name”), with the hope of “shaming” it and changing its course of action. By randomly distributing the treatment we are able to control for various causes which may effect the likelihood of a particular decision maker to adopt a norm-conforming policy (such as ideology, ethnicity, gender or age).

4For example: Amnesty International report on Israeli policies towards Palestinian prisoners
4.1.2 Neutral Information (I)

In order to reject the notion that new information, not emotions, has led to policy change there is a need to account for I as a separate independent variable. To do so, we can randomly assign a third group a text which includes neutral information about the same topic. For example, the Amnesty International report includes normative adjectives such as “cruel”, “ruthless” or “devastating”, eliminating these words may result in a more neutral and factual text. In taking these words out we can create a baseline text which is constant across treatments. Thus, if we find that exposure to such a text is less likely to cause leaders to change their policy, we can rule out this alternative explanation.

4.1.3 Numeric Indicators (N)

The third explanation asserts that elites are not driven by emotions or new information, but rather by their wish to maintain their state’s reputation. Since reputation is not absolute, but rather relative to other states, we could operationalize it by adding numeric indicators and rankings to a third distributed text (R). Comparing publicized countries to other states initiates a stronger sense of competitiveness and a fear towards the state’s placement in relation to others, thus hurting its reputation (Kelley and Simmons 2015). If decision makers believe that Israel is amongst the highest violators of human rights, they may be driven to change their policy preference. We therefore add these numeric indicators to the baseline text. Once again, if we find that exposure to such a text is less likely to lead to policy preference change we would be able to rule out this alternative explanation.

4.2 Moderator (M)

Since the proposed theoretical framework suggests that emotions of guilt and shame are moderated by political self-efficacy, there is a need to account for this fourth variable as well. This variable will be conceptualized as the degree to which actors believe in their own capabilities to control political change (M). Following the works of others, I operationalize political self-efficacy with a questionnaire (forming an index variable) that has been frequently used and already tested on political elites (Caprara et al. 2009b). Since the groups are randomly assigned into treatment and control we should have similarly distributed levels of political self-efficacy between the treatment and control. However, it may be more useful to account for this variable before subjects are asked to read the text in order to ensure that levels of political self-efficacy are not altered by the treatment.

4.3 Dependent variable

The dependent variable, policy preference change (Y), is defined as the adoption of a new course or principle of action in accordance with the violated international norm. In or-
der to account for this variable, data will be collected in a two stage experiment. In the first stage, elites will be exposed to relevant texts, as outlined in the previous section. In the second stage, one week later, leaders will be asked to sign a petition which supports a change in policy in accordance with the violated norm. In this scenario, we would ask respondents to sign a petition calling to change Israel’s policy so that Palestinian prisoners will be held in the occupied territories and allowed family visits. Although this measurement does not capture an actual change in policy, it allows us to measure more sincere preferences and minimize concerns of “cheap talk” and mere intentions.

4.4 Respondents

As mentioned, the variables outlined above will be accounted for in an elite survey experiment. Subjects will be recruited from Israel’s Ministry of Foreign Affairs’ cadets course. These are individuals who undergo a six-month training before integrating into a variety of policy roles within the ministry of foreign affairs. Access to the course will be gained through a contact who has completed the course and can distribute surveys to current and future cohorts of the course.

In surveying political elites in training we are able to overcome two potential criticisms. First, if we were to survey the general public a valid criticism may be that those who go into policy have certain characteristics that differ from the average citizen such as high socio-economic status, education, charisma etc. The individuals who undergo this course have been carefully selected (only 1% of applicants are accepted), and have gone through psychological, academic and security screening which evaluated them as fit for a role in international politics. We can therefore assume that individuals accepted to this course acquire the traits of decision makers and other political elites.

Second, one may argue that decision makers are repeatedly faced with criticism of Israel’s policies and constantly read reports similar to the ones distributed by Amnesty International. If this is true, both our control and treatment groups have been treated before the experiment. While we cannot completely solve this issue, we are able to minimize it by targeting elites in training. Since these individuals did not begin their position yet, we conjecture that they have been exposed to less criticism than those who have been in office for a long time.

5 Potential Limitations

Despite the ability of the research design described above to provide insight into the causal effects of “naming”, it is subject to multiple limitations. In this section, I elaborate these challenges and address potential ways in which they can be addressed and minimized.

---

5Israel’s Ministry of Foreign Affairs’ Cadets Course
Nevertheless, some of these issues could not be solved in the scope of this research and call for further investigation which will be discussed in the concluding section.

5.1 Attrition

One potential limitation may be selective attrition. If a large group of respondents drop out of the study or do not answer some of the questions, it would be rather difficult to make inference. When sampling elites, it is often a concern that individuals would not want to participate due to busy schedules. However, since I intend to conduct the experiment within the scope of the course, it is unlikely that this would play a role. Nevertheless, in cases of many missing values it is essential to understand why attrition takes place. In other words, are respondents dropping out randomly, or because of specific factors that have to do with the research design (ie. measurement or treatment)? If it is the first case, we may be able to correct our results by simply collecting more data by repeating the experiment. However, if attrition is a result of flawed research design, our measurements should be re-evaluated. In this case, we may want to rephrase the texts distributed as a treatment, or modify the measurement of policy change. For example, some individuals may be reluctant to sign a petition since they do not want to share personal information (such as name) with an unknown factor. I would therefore have to re-assess my measurement of policy and come up with an anonymous one that captures the same phenomenon.

5.2 Limited Access

Elite samples are often rather small as they represent a small fraction of the society. The design I described above could be understood as an ideal-type experiment, as it tests for all possible mechanisms within this scenario. However, in reality, I may not be able to assign three different treatment groups due to limited access to elite samples. It may be preferable to select two treatments which are most likely to serve as explanatory variables and address the third in another way, such as a manipulation check. For example, if we decide to implement the emotion and reputation treatments, we could ask all respondents after the experiment whether they knew of Israel’s policies towards Palestinian prisoners. Another way in which this issue may be addressed is by conducting a pilot experiment in order to evaluate the feasibility of the project and improve the design accordingly. Since it is a concern to abuse connections with elite sampling, it may be preferable to conduct this pilot on the general public, taking into account differences between the two samples when drawing conclusions.
5.3 Alternative Mechanisms

In this project I design a research that tests the direct effect of “naming” on policy change. However, constructivist theory suggests that there are other, indirect ways, in which this tactic could promote international norms within the socialization process. According to the theory, publicizing states’ violations of international norms could promote policy change in three other ways as well. First, publications could provide incriminatory information to other countries and transnational actors which pressure the accused government from above (Katzenstein 2013). For example, the Amnesty International report could inspire US officials to pressure Israel into conforming with the International norm, thus leading to an indirect effect. Second publications could lead to pressure from below, through public opinion. Leaders themselves may not be inclined to change policy following a publication, but the public may pressure them to do so for various reasons. Third, naming may have a spillover effect, resulting in changes in less related domains. For example, Israel may not want to change its policy with regards to Palestinian prisoners but would improve its behavior in some other human rights issue instead.

These three mechanisms provide fascinating theories to the ways in which socialization occurs at the international level. However, they are outside the scope of this research and will not be able to be accounted for. Future research should strive to test these causal relationships. Similar strategies to those offered within this design could be applied. For example, US decision makers could be asked to sign a petition that deals with US relations with Israel, pressuring it to conform with international laws. Another possible research could account for public support through a similar experiment within the Israeli public.
References


Kelley, Judith G and Beth A Simmons. 2015. “Politics by number: Indicators as social pressure in international relations.” American journal of political science 59(1):55–70.


Appendix

Optional Texts (Based on Amnesty International Reports):

Shame/Guilt text:

“Israel’s ruthless decades-long policy of detaining Palestinians from the occupied West Bank and Gaza in prisons inside Israel and depriving them of regular family visits is not only cruel but also a blatant violation of international law. Devastating testimonies from family members and Palestinian prisoners detained in the Israeli prison system shed light on the sorrow and suffering endured by families who in some cases have been deprived from seeing their detained loved ones for many years. Some of these individuals are young children, under 13 years of age, who have been denied the right to see their parents for many years. Under international humanitarian law, detainees from occupied territories must be detained in the occupied territory, not in the territory of the occupying power.”

Information text:

“Israel’s decades-long policy of detaining Palestinians from the West Bank and Gaza in prisons inside Israel and depriving them of regular family visits is a violation of international law. Palestinian prisoners detained in the Israeli prison system have been deprived from seeing their families for many years. Some of these individuals are young children, under 13 years of age, who have been denied the right to see their parents for many years. Under international humanitarian law, detainees from occupied territories must be detained in the occupied territory, not in the territory of the occupying power.”

Numeric indicators text:

“Israel’s decades-long policy of detaining Palestinians from the West Bank and Gaza in prisons inside Israel and depriving them of regular family visits is a violation of international law. Palestinian prisoners detained in the Israeli prison system have been deprived from seeing their families for many years. Some of these individuals are young children, under 13 years of age, who have been denied the right to see their parents for many years. Under international humanitarian law, detainees from occupied territories must be detained in the occupied territory, not in the territory of the occupying power. Recently, Israel was the only western democracy ranked as one of the 38 countries in which human rights are frequently abused. Global indicators frequently place Israel amongst the highest violators of international law, first amongst OECD members.”