Shirking and Signaling: Avoiding Domestic Reputation Costs in Multilateral Institutions*

Julia C. Morse† and Tyler Pratt‡

October 29, 2018

Abstract

Reputation underpins many theories of international cooperation, but few studies have rigorously tested how and why international organizations (IOs) affect reputation. Given that governments often vigorously challenge allegations of non-compliance, under what conditions do IOs cut through the noise? We test the relationship between IOs, government reputation, and domestic audiences in the issue area of human rights. Human rights scholars argue that signals of non-compliance can reduce domestic support for leaders and mobilize opposition. Governments, however, can use various "shirking" strategies — from citing national security concerns to attacking the objectivity of monitoring — to disrupt such signals. We use a survey experiment to analyze how these rhetorical challenges interact with information about non-compliance to shape government approval and respondents' willingness to donate to human rights activists. Our preliminary results suggest that shirking works – governments can reduce the reputation cost of non-compliance by adopting certain public communications strategies – but only when monitoring is performed by a group of member states or a non-governmental organization. In contrast, IO monitoring by independent experts is able to withstand such attempts. This finding suggests a key function of multilateral institutions is not only providing information, but also making information more resistant to rhetorical challenges.

---

*Authors' Note: An earlier draft of this paper was presented at the Southern California Methods Conference in September 2018.

†Assistant Professor, University of California, Santa Barbara. Email: jcmorse@polsci.ucsb.edu

‡Assistant Professor, Yale University. Email: tyler.pratt@yale.edu

1 Introduction

In the early 2000s, the United States government engaged a prolonged campaign to dispute allegations that it tortured terrorist suspects in violation of domestic and international law. In a now-infamous memo, the US Justice Department argued that President Bush could declare the Taliban and Al Qaeda outside the coverage of the Geneva Conventions and other international treaties prohibiting torture. In another memo, the Office of Legal Counsel reinterpreted existing law, arguing that a wide array of interrogation practices were permissible because they did not meet the precise definition of torture included in the 1984 Convention Against Torture (CAT) (Sanders, 2011). This definition conveniently precluded acts such as confinement in small spaces, sleep deprivation, and waterboarding.

Domestic and international audiences largely rejected the Bush administration's logic. Throughout the administration's eight years in office, a majority of Americans remained opposed to the use of torture or enhanced interrogation techniques (Gronke *et al.* , 2010). In 2005, the United Nations Committee Against Torture accused the United States of violating a long list of CAT provisions, including the physical abuse of detainees and the rendition of suspects to countries where they faced a risk of torture (Committee Against Torture, 2006). In 2006, the US Supreme Court ruled that all detainees be granted protections afforded by the Geneva Conventions. Despite these decisions, former Bush administration officials have remained committed to the facade of compliance with rules banning torture. In a 2014 interview, former Vice President Dick Cheney offered such a defense: "There's this notion that somehow there's moral equivalence between what the terrorists do and what we do. And that's absolutely not true. We were very careful to stop short of torture" (Bradner, 2014).

The Bush administration's stalwart defense illustrates the degree to which governments seek to manage their domestic and international reputations. In public, most leaders deliver

carefully worded statements and hold targeted press conferences. In private, they defend questionable policies to allies and adversaries alike. Perhaps for this reason, a large body of literature in international relations posits that reputation is key to understanding both conflict and cooperation. In the conflict arena, states cultivate reputations for resolve to deter adversaries and to increase bargaining leverage (Schelling, 1960, 1966; Mercer, 1996; Huth, 1997). In the realm of cooperation, fostering a reputation for compliance provides states with greater returns in repeated patterns of interaction. States that are tempted to violate a commitment may opt to comply because they fear being excluded from future cooperation (Keohane, 1984; Guzman, 2008).

The logic of reputation is particularly attractive for students of multilateral cooperation. International organizations (IOs) typically lack formal enforcement tools to force states to follow their rules. Reputation offers a separate mechanism through which IOs can impost costs on rule violators. Simply by broadcasting a signal of non-compliance to domestic or international audiences, IOs can trigger costly changes in a state's reputation. Other governments may respond to the IO signal by curbing ties with the non-compliant state, while domestic actors may punish the government or mobilize to promote policy change.

Despite its privileged status among scholars of international cooperation, however, there are reasons to question whether and how reputational mechanisms affect outcomes. Reputation is multifaceted – each country maintains dozens of reputations (Downs & Jones, 2002), reflecting different traits, behavioral tendencies, and behaviors (Dafoe *et al.* , 2014). Even if IO monitoring can successfully reduce a state's reputation for compliance, the effect is likely to be muted as domestic and international audiences weigh this particular dimension of reputation against others. Moreover, governments routinely employ strategies to contest IO signals of non-compliance. They point to extenuating circumstances to argue that an alleged violation does not reflect the government's underlying propensity for compliance. At other times, they attack the objectivity or legitimacy of the IO itself. These strategies are

designed to shirk reputation costs by manipulating the inferences drawn by relevant audiences. Given these caveats, the importance of reputation as a causal mechanism is likely to vary significantly across issue areas and policy domains.

This paper probes the role of reputation in international cooperation on human rights. Cooperation in this issue area is highly institutionalized; most states have signed on to multiple international human rights treaties. While IOs that focus on human rights have few enforcement powers, the United Nations includes ten human rights treaty bodies that monitor and publish reports on the implementation of the core treaties. These bodies produce significant information about violating states. Monitoring reports are made public, in part, to trigger the reputational mechanisms discussed above.

States have particularly strong reasons to care about their domestic reputations in this issue area.[1] Existing scholarship suggests international human rights agreements may shape state conduct through domestic mobilization (Simmons, 2009) and through the internalization of norms (Hafner-Burton & Tsutsui, 2005; Risse & Sikkink, 1999).[2] If an IO publishes a report about a government's human rights violations, such reporting could exact costs on a government by decreasing public support or increasing anti-government mobilization. Like all informational transmissions, however, declarations about non-compliance contain a mixture of signal and noise. Countries will work to mute an IO's signal of non-compliance so that it inflicts less reputational damage on the government.

We examine three common strategies that countries use to shirk the reputation costs of non-compliance: national security challenges, bias challenges, and sovereignty challenges. Governments frequently use national security considerations to justify violations of human

[1]Because human rights agreements regulate domestic conduct that produces few externalities for other states, a state's international reputation on human rights is likely to be less influential in this context than in others (Simmons, 2009).

[2]Notably, the effect of human rights agreements on domestic policy may be contingent upon a state's level of democracy and civil society (Neumayer, 2005).

rights, particularly when the rights in question pertain to suspected terrorists. Alternatively, a government may challenge the objectivity of a monitoring body by claiming the body is biased against the country. Finally, a government may attack a monitoring institution by arguing it is interfering with the affairs of a sovereign nation. While each strategy adopts a slightly different approach to protecting the government's reputation, all of them attempt to re-contextualize IO monitoring in a way that limits the inferences of third parties.

We argue that, in general, shirking strategies can successfully reduce the reputation costs of non-compliance with human rights commitments; however, certain types of international monitoring bodies may be able to reduce the effectiveness of shirking. We hypothesize that IO monitoring from independent experts is more resistant to shirking than IO monitoring produced by member states, and even compared to monitoring from a non-governmental organization (NGO). Domestic audiences are more likely to view IO expert bodies as credible and independent compared to an IO panel of member states; these perceptions of independence should make an IO more resistant to shirking. While NGOs may also be viewed as credible monitors, we expect that respondents may be more sympathetic to government shirking strategies because unlike in an IO, the government did not consent to be monitored in the first place.

We probe the effect of shirking strategies on reputation through a survey experiment completed by 1,448 respondents from Amazon Mechanical Turk (MTurk). The respondents were presented with a scenario in which the US government is accused of violating the Convention Against Torture.[3] We show that shirking strategies based on national security or sovereignty challenges successfully mitigate the reputation cost of non-compliance; bias challenges, however, have little effect. In addition, we find compelling evidence that IOs have a unique ability to resist shirking, but only when IO monitoring is done by a panel of

[3]We plan to supplement this data by repeating the survey experiment on a nationally representative sample in the United States.

independent experts.

Our findings have several implications for research on international cooperation. First, we add to the body of scholarship underscoring the constraints of reputation as mechanism for sustaining cooperation. While existing work emphasizes the potential for states to hold multiple, distinct reputations, we explain how strategic behavior by non-compliant states can deflect reputational damage away from the state and toward other actors or institutions. Our findings suggest both NGOs and certain types of international institutions may be constrained in the degree to which they can rely on reputational mechanisms to ensure compliance. By highlighting the significance of context for reputational effects, our research may also have implications for scholarship that relies on reputation as a key causal mechanism. Literature on ratings and ranking (Cooley & Snyder, 2015) and global performance indicators (Kelley & Simmons, 2015, 2018), for example, identifies reputational damage as a primary way that such information affects state behavior. Our findings suggest one reason why reputation might matter in these contexts: perhaps rating and ranking states is effective not just because transmits information, but also because it resists shirking behavior by governments.

Second, we highlight the significance of institutional design in shaping the behavior of member states. Our empirical results indicate that credible, independent IO monitoring bodies are more immune to shirking than monitoring by either IO member states or NGOs. This finding is consistent with work emphasizing the benefits of delegation (Hawkins *et al.* , 2006). Although states must cede control when delegating monitoring to international bureaucrats, they gain an institution that is more robust to attacks from non-compliant actors.

2 Reputation Costs and International Cooperation

Scholars of international relations have long recognized the potential for reputational effects to sustain international cooperation. International politics occurs amidst great uncertainty about the capability and intentions of other actors. In the presence of such uncertainty, states and non-state actors make decisions about how to interact with others based on reputation – that is, a belief or judgment about an actor's trait, type, or behavioral tendency (Dafoe *et al.* , 2014, 365). A key insight from this definition is that beliefs, not facts, constitute the basis of reputation. Information about a state's behavior may influence its reputation, but ultimately, reputation depends on observers integrating new information in a way that ascribes a state's behavior to underlying traits or behavioral tendencies (Mercer, 1996, 6).

Because a country's reputation is defined by the beliefs of outside observers, recent work in international relations has sought to probe this relationship through survey experiments. This trend has been particularly prominent in work on audience costs. Early research posited that democratic leaders tie their own hands when they publicly threaten another state because domestic constituencies will punish governments that are inconsistent (Fearon, 1994; Baum, 2004; Slantchev, 2006; Weeks, 2008). Reputation as a causal mechanism is thus relevant at two points: a leader's international image and his or her domestic image. Survey experiments have probed the latter part of this mechanism, examining the scope and logic of audience costs in numerous contexts (e.g. (Tomz, 2007; Trager & Vavreck, 2011; Levendusky & Horowitz, 2012; Davies & Johns, 2013; Kertzer & Brutger, 2016; Brutger & Kertzer, 2018)). While empirical support for audience costs is strong, this relationship can be mitigated by preexisting policy preferences (Chaudoin, 2014) or framing effects (Levendusky & Horowitz, 2012).

If beliefs and uncertainty affect a state's propensity for conflict, they are also likely to affect the propensity of states to cooperate. Keohane (1984) hypothesizes that uncertainty

is a key impediment to cooperation. By stabilizing expectations and clearly identifying proscribed behavior, IOs facilitate the development of reputations, which creates incentives for states to engage in cooperative behavior. As Keohane argues, "a good reputation makes it easier for a government to enter into advantageous agreements; tarnishing that reputation imposes costs by making agreements more difficult to reach" (Keohane, 1984, 105-106). Reputation is one way that states can help resolve information asymmetries that might otherwise preclude mutually beneficial cooperation among states.

2.1 Reputation as a Driver of Compliance

Scholars of international organizations consider reputation to be a particularly powerful force for sustaining compliance.[4] If states interact over multiple time periods, the reputation that they accrue in earlier periods will influence their success later on. When a state is observed violating an international agreement, other actors update their beliefs about the state's likelihood of complying in the future. They can then deny the violator the benefits of future cooperation by screening them out of future agreements.

Scholars have found empirical support for the logic of reputational effects in a variety of contexts. Axelrod (1984) uses a series of simulated cooperative interactions to show how actors can maximize their own payoffs by conditioning on the past behavior of cooperative partners. Milgrom *et al.* (1990) demonstrate the importance of reputation, as well as the institutions that bolster it, in enabling trade among merchants in medieval Europe. Other scholarship highlights the role of reputation in facilitating compliance with sovereign loan agreements (Tomz, 2007), monetary rules (Simmons, 2000), trade agreements (Kono, 2007), and the laws of war (Morrow, 2007).

Many international organizations have public monitoring schemes that implicitly rely on reputation to incentivize cooperation. IO monitoring schemes range from voluntary state re-

[4]See Martin & Simmons (2012, 337) for a more detailed discussion of this point.

porting to independently investigating and publicizing instances of non-compliance (Bradley & Kelley, 2008). The United Nations Security Council, for example, monitors the implementation of targeted sanctions against Al-Qaida, but relies on states to self-report on specific laws and procedures for freezing terrorist assets. In contrast, the Financial Action Task Force, a technocratic body that also focuses on terrorist financing, collects detailed information and travels to monitored countries to assess compliance. Whereas the Security Council discusses only general trends in behavior, the Financial Action Task Force publishes a non-complier list that publicly identifies countries with deficient policies.

IO monitoring is particularly effective at signaling information to domestic and international audiences when it is independent and precise. When IO monitoring is delegated to a secretariat or to bureaucrats, it is more likely to be a credible source of information because it is insulated from interstate politics. Although such acts limit state control, member states are often willing to make this tradeoff to take advantage of technical expertise (Hawkins *et al.* , 2006). IO monitoring will also be more likely to impact a state's reputation when the monitoring report sends a clear and precise signal about state behavior. Precise monitoring, such as through a global ranking, blacklist, or performance indicator, reduces uncertainty and facilitates comparisons across countries (Kelley & Simmons, 2015; Cooley & Snyder, 2015; Kelley & Simmons, 2018). Finally, the impact of IO monitoring on a country's reputation will depend on the salience of the information to the audience. Even a credible, precise allegation of non-compliance may not significantly change an audience's assessment of a government if the audience places a low priority on the government's propensity to comply. If, on the other hand, IO information is relevant to a group's priorities, the group may change its behavior toward the monitored state based on the new information, imposing reputation costs.

2.2 Reputation Costs

When states, non-state actors, or domestic audiences update their beliefs about a state's type, reputational damage will only drive compliance if these actors also change their behavior toward the non-compliant state. A reputation cost occurs when actors respond to new information by voluntarily adjusting their behavior toward a non-compliant state, reducing its current or future welfare. This change in behavior can occur for two reasons. First, as previously discussed, observers may adjust their behavior because an IO signal provides information about the non-compliant state's underlying traits or characteristics. Second, observers may adjust behavior because the informational signal provides information about how other observers are likely to view and behave toward the non-compliant state.[5]

3 Shirking Reputation Costs

As the preceding discussion makes clear, IO signals about state compliance operate in a highly complex environment. Consider the unbroken chain of events that must occur for reputational effects to incentivize compliance. First, IOs or other actors must identify violations of international rules. In some cases, states authorize IOs to operate as information collectors and assessors, conducting regularized reviews of state policy. In other cases, IOs are dependent upon states or non-state actors to highlight instances of non-compliance. Victims of non-compliance themselves may even be key sources of information (Dai, 2007).

Once an IO determines a state is failing to fulfill its obligations, it must disseminate information about non-compliance to actors that can impose costs. Recent scholarship sug-

[5]Dafoe *et al.* (2014, 374) term this distinction "first-order and "second-order" beliefs. First-order beliefs reflect an actor's observations about another actor's characteristics or behavioral tendencies. Second-order beliefs are an actor's beliefs about what a larger group of observers believes. While most work on reputation focuses on first-order beliefs, the decision to censure or punish rule violators often depends significantly on second-order beliefs. For example, permanent members of the United Nations Security Council may prefer to avoid vetoing a decision punishing an instance of non-compliance if their position is isolated.

gests the specific form of information transmission matters – information spread through ratings, rankings, indices, or blacklists may increase reputation costs because it facilitates comparisons across states and reduces uncertainty (Kelley & Simmons, 2015; Cooley & Snyder, 2015; Kelley, 2017; Morse, 2017). Other IOs signal information about non-compliance through public statements and reports. The UN Human Rights Committee, for example, issues a public "list of issues" identifying suspected violations of the International Covenant on Civil and Political Rights. States then have an opportunity to contest these findings before a final judgment is made.

Finally, an audience must interpret an IO's informational signal. When information damages a country's reputation, audiences will update their beliefs about the non-compliant government (or, in some cases, the IO) and potentially adjust their behavior. Both citizens and other governments are likely to differ in how they respond to such information. Staunch supporters may remain steadfast, viewing the IO as biased rather than imposing any costs on the non-compliant government. But others may update their beliefs about the reliability, character, or competence of the government. For example, citizens may view the government as more likely to violate additional international or domestic commitments in the future; alternatively, they may interpret the IO signal as evidence that the government has engaged in particularly egregious behavior worthy of international censure.

This final step, in which actors interpret a signal of non-compliance and adjust their behavior, has been under-theorized in the existing literature. Existing work neglects the fact that non-compliant states have an opportunity to shape how audiences perceive an alleged rule violation. Moreover, even when citizens impose reputation costs on non-compliant states or become activists for increased compliance, the micro-foundations of this process remain unclear. Do citizens engage because they view the government as less trustworthy? Do they fear that similar violations could put them at risk of physical harm? It is at this stage in the process – after an IO has publicly identified non-compliant behavior – that we focus our

analysis.

When an IO signals that a state is non-compliant, the state can respond in one of several ways. At one end of the spectrum is complete acceptance – the state can accept the description of its behavior and any concomitant reputation costs likely to be imposed. Often, states that adopt this strategy acknowledge the validity of the monitoring process and promise future improvements in compliance. At the other end of the spectrum, a state might reject an IO's jurisdiction completely, perhaps even opting to exit the institution. For example, after the United States lost two cases in the International Court of Justice (ICJ) over violations to the Vienna Convention on Consular Relations, the United States withdrew from the Optional Protocol that granted the ICJ jurisdiction over this issue (Quigley, 2009). In the middle of these two extremes, states opt for strategies that seek to re-contextualize informational signals and mitigate reputation costs. These are the strategies that we focus on in this paper.

3.1 Shaping Signals and Interpretations

Efforts to shape the interpretation of non-compliance signals are important because drawing inferences from rule violations is not straightforward. A violation might be a signal of a state's general unwillingness to abide by international obligations. It could also signal a lack of capacity to implement the rule, or a misunderstanding regarding the rule's applicability in a specific context. While IOs would prefer that states interpret non-compliance in a way that maximizes reputation costs, states that seek to flout international rules have the opposite preference. Strategic maneuvering by non-compliant states does not stop once monitoring has uncovered problematic behavior.

Mercer (1996) highlights the important role of interpretation in shaping reputational effects. For information to change a state's reputation for compliance, observers must decide the state's behavior is attributable to dispositional rather than situational characteristics.

Despite the focus in the existing literature on dispositional traits, observers often consider situational factors when drawing inferences about non-compliance. Since 2009, for example, the International Monetary Fund (IMF) has surveyed and analyzed public finance across a host of different countries through its "Fiscal Monitor" report. In 2009 and 2010, states and non-state actors reading the Fiscal Monitor would have been much more likely to attribute negative reports about a state's fiscal and macroeconomic health to the recent financial crisis. In contrast, if a country receives a negative review in 2015, audiences may be more likely to attribute poor performance to government mismanagement. When observers update their beliefs about a state's characteristics based on IO monitoring, IO-produced information has directly affected a country's reputation.

Uncertainty over the appropriate interpretation of rule violations provides space for contestation. Non-compliant states can mitigate reputation costs by successfully obscuring informational signals from IOs. After an IO publishes a monitoring report or publicly identifies a rule violation, the accused state may re-contextualize or even openly challenge these findings. The goal of this behavior is to alter how other states or non-state actors interpret the accused state's violation. If successful, actors will not draw strong inferences about the accused state's commitment to international rules. Instead, they will form updated beliefs about the monitoring institution or the circumstances surrounding the alleged violation.

In the following sections, we highlight three strategies that states often use to undermine domestic reputation costs. We call these strategies *national security challenges*, *bias challenges*, and *sovereignty challenges*, depending on the principle they invoke to contest allegations of non-compliance. These strategies are employed after an international institution issues a public judgment that a state has violated international rules.[6] When a state

[6]Public judgments may take the form of monitoring reports, regular reviews of state behavior, statements from international bureaucrats, or ruling from international courts. We distinguish between arguments offered by states as part of a review process (e.g., international court proceedings) and shirking strategies they employ after they have been deemed non-compliant. The former are attempts to avoid a judgment of non-compliance, not strategies to alter reputational effects.

launches a national security challenge, it claims that the behavior in question was necessitated by an extreme security crisis. This challenge aims to convince an audience that the violation stemmed from specific extenuating circumstances, meaning the audience should adjust its beliefs about the situation that spurred the violation, not the disposition of the government. National security challenges may also serve to increase the salience of a second type of reputation: a government's ability to protect its citizens. Alternatively, a bias challenge calls into question the credibility of an IO, its rules, or the monitoring process. Governments encourage the audience to interpret the alleged violation as evidence of unfair treatment by the IO. Finally, a sovereignty challenge contests the legitimacy of the IO to govern and monitor the state's behavior. This challenge appeals to an alternative vision of world order where sovereign states are solely accountable to their citizens, not international rules or norms. It asks the domestic public to interpret the IO monitoring report as a signal of global governance institutions run amok.

3.2 National Security Challenges

A national security challenge occurs when a non-compliant government argues its transgression was a necessary response to a security crisis. In other words, a government does not contest the accuracy of IO monitoring but rather argues that the situation demanded a violation of international rules. States that use this strategy claim that the violation does not reflect the underlying disposition of the government and should not be used to draw inferences about future behavior (especially once the security crisis has passed). This strategy has been particularly pronounced in recent years, as governments adopt extraordinary measures to deal with suspected terrorists. The United Kingdom (UK), for example, passed several controversial anti-terrorism laws in the period immediately following 9/11. UK Home Secretary David Blunkett justified anti-democratic measures, such as holding foreign nationals for indefinite periods without trial, by comparing the UK's susceptibility to attack to the

emergency situation faced by Britain during World War II (Tempest, 2004).

National security challenges, as well as other similar appeals to extenuating circumstances, exploit the fact that audiences must simultaneously manage two sources of uncertainty when interpreting IO signals. First, they are uncertain of the government's general propensity to abide by its international commitments. In the words of Guzman (2008), audiences don't know the "reputational payoffs" of a government — i.e., how much the government prioritizes compliance with its commitments. In making the decision to comply or violate, however, a government must weigh these reputational payoffs against all other interests affected by the decision. These "nonreputational payoffs" are also unknown to audiences. Confronted with an IO signal of non-compliance, an audience may update its beliefs about either the government's general willingness to comply (reputational payoffs); the government's other interests connected to the decisions (nonreputational payoffs); or some combination of the two. National security challenges represent a claim that nonreputational payoffs overwhelmingly favored violation.[7] If they are believed, rational audiences will conclude that even a government that placed a high value on honoring its commitments would have violated the relevant rule.

3.3 Bias Challenges

Bias challenges occur when a non-compliant state attempts to weaken reputation costs by criticizing the credibility of an IO, its rules, or its monitoring process. In these challenges, a state undermines the accusation of non-compliance by arguing that the monitoring process is biased against the violating state. The goal of a bias challenge is to convince a domestic audience to react to signals of non-compliance by drawing negative inferences about the institution, not the accused state.

[7]Guzman (2008) makes this very point, arguing that a violation "that is plausibly attributable to the fact that the country is at war and devotes its efforts to the pursuit of the war effort rather than compliance...will generate a smaller reputational sanction than a violation that cannot be justified in some similar way" (78).

The logic of a bias challenge relies on the assumption that more objective monitoring generates higher reputation costs. A significant body of scholarship links objectivity and procedural legitimacy with better compliance behavior. International legal scholars argue that the legitimacy of norms and rules affects the probability that states will comply with them (Franck, 1990; Koh, 1998). Rules created by IOs are more likely to be perceived as fair than those created by states or non-state actors, because IOs embody a type of "rational-legal authority that modernity views as particularly legitimate and good" (Barnett & Finnemore, 1999, 707).

Bias challenges contest the accuracy of signals from IOs. States challenge the content of monitoring directly, claiming the information itself is wrong. The immediate goal of such a descriptive claim is to eliminate reputation costs by convincing other states that the information is false. For such a strategy to succeed, however, a non-compliant state usually needs to attack the competence of the IO itself to explain why the monitoring report is incorrect. For example, in 2015 the World Anti-Doping Agency (WADA) accused the Russian government of perpetrating a systematic, state-sponsored doping program that delivered performance-enhancing drugs to the country's athletes. Russian officials strongly denied the accusations, calling them "baseless" and "fictional" (Holdsworth, 2015). They simultaneously attacked the integrity of institutions decision-making process, arguing that the WADA chairman overstepped his authority to single out Russia at the behest of Western states (Stinson, 2016).

3.4 Sovereignty Challenges

Sovereignty challenges occur when governments dispute the right of the IO to set constraints on the state's behavior. Unlike a bias challenge, where governments claim an IO has failed to carry out its mission in a fair and credible manner, sovereignty challenges contest the legitimacy of the IO as an actor. These challenges cite the principle of state sovereignty to

argue that external actors have no right to dictate the domestic policies of national governments. They often highlight the lack of accountability in international institutions and organizations compared to national governments, which are answerable to citizens. When Ecuador, for example, found itself on a "blacklist" of states due its weak oversight of money laundering and terrorist financing, the government attacked the legitimacy of the institution, not the information. Ecuadorian Foreign Minister Ricardo Patino stated that the Ecuadorian government did not recognize the right of the monitor (in this case, a 35-country intergovernmental body) to "dictate policy" (The Andean Laundry, 2010).

When a government launches a sovereignty challenge, it asks audiences to interpret the IO signal as evidence of an illegitimate infringement on national sovereignty. If the challenge is successful, actors primarily update their beliefs about the monitoring institution and not the propensity of the government to honor its commitments. By invoking broad principles of accountability and legitimacy, sovereignty challenges may also succeed by re-framing accusations of non-compliance. Governments that might lose domestic approval in a narrow dispute over human rights violations, for instance, could increase its approval by shifting to a broader dispute over national sovereignty. In that case, sovereignty challenges function by increasing the salience of other dimensions of a government's reputation (e.g., its ability to guard citizens from foreign influence).

Because sovereignty challenges strike themes that are often associated with nationalist or populist leaders, the current trend towards populism may increase the prevalence of such challenges. However, sovereignty challenges entail several consequences that might limit their attractiveness to states. First, they do not completely blunt an IO's signal about the tendencies of the accused state to respect international commitments. In fact, they risk reinforcing the signal by rejecting international constraints on the government's behavior. Second, successful sovereignty challenges partially absolve the accused state by undermining the legitimacy of the monitoring IO. Audiences update their beliefs about the institution,

which may threaten valuable cooperation in the future. If non-compliant states experience benefits from cooperation in the institution, they may be hesitant to attack its legitimacy to justify a violation of its rules.[8]

3.5 Theoretical Expectations

We expect that in general, all three shirking strategies will mute the reputation costs of violating international rules. In the survey experiment described below, we measure reputation costs in two ways. The first is the change in respondents' expressed support for a government accused of violating its commitments. Second, we measure whether respondents are willing to engage in costly action in the form of donating their survey proceeds to a pro-compliance advocacy group.

We examine whether each shirking strategy changes respondents' tendency to support the government and take costly action, compared to the control condition in which the government offers no defense of its behavior. Under the assumption that information about a government's human rights violations is more likely to decrease than increase public support, we interpret *increases* in domestic approval in the shirking conditions as a *decrease* in reputation costs. In other words, we expect respondents are more likely to approve of the government when it justifies its behavior than when it offers no defense. However, we hypothesize that when information about non-compliance comes from an IO with independent experts, shirking will cause a smaller reduction in reputation costs.

- *H1: Shirking Strategies - When a government is accused of non-compliance by an IO, the use of national security, bias, and sovereignty challenges will increase support for*

[8]For example, when the World Trade Organization (WTO) ruled in 2003 that the Bush Administration's imposition of steel tariffs violated its rules, U.S. officials refrained from directly attacking the legitimacy of the institution. As a major beneficiary of cooperation in the WTO, the United States would have experienced significant costs if a sovereignty challenge were successful. Instead, the Administration decided to rescind the tariffs in a tacit admission of the violation.

the government compared to an uncontested posture.

- *H2: Identity of Monitor and Shirking - National security, bias, and sovereignty challenges will be less effective when IO monitoring is conducted by independent experts compared to member states or a non-governmental organization.*

4 Data and Results

4.1 Survey Experiment

As an initial test of the effect of these shirking strategies, we conducted an internet-based survey experiment in October 2018 using Amazon Mechanical Turk (MTurk). A total of 1,448 U.S.-based respondents participated in the experiment.[9] Like other respondent pools recruited through MTurk, our sample is not nationally representative (Berinsky *et al.* , 2012).[10] However, comparisons between MTurk and nationally representative samples suggest that the political and psychological dispositions of MTurk respondents closely mirror those of the mass public, mitigating concerns about external validity (Clifford *et al.* , 2015).

Respondents were asked to complete a survey about foreign affairs and U.S. foreign policy.[11] Upon entering the survey, each respondent was presented with one of two brief descriptions of international human rights rules regarding extradition to countries that practice torture. One description drew attention to an international institution (the Convention Against Torture) and an associated monitoring body, and the other emphasized a non-

[9]We restricted the respondent pool to MTurk workers who were based in the United States, who had completed at least 100 MTurk tasks, and who had a record of completing tasks at a rate of at least 90%.

[10]Specifically, the sample skews male (58.1%), white (80.8%), and educated (61.8% with college degree) compared to the national average. Slightly over half of respondents (51%) either identify as a democrat or lean towards the democratic party; the corresponding proportion of republicans and independents are 35.5% and 13.5%, respectively. The modal age range is 25-34, though nearly half (48%) of the sample is at last 35 years old and 21.2% are 55 or older.

[11]The complete survey is included in the appendix.

governmental organization (Human Rights Watch).

1. The United States is a member of the Convention Against Torture, an international treaty that seeks to promote justice and human rights around the world. Under this agreement, governments commit that they will not allow any form of torture or excessive use of force against people in their countries. They also agree not to transfer individuals to the custody of other countries where they are likely to be tortured. An international body affiliated with the Convention monitors compliance with human rights rules.

2. Human Rights Watch is a non-profit, non-governmental organization that seeks to promote justice and human rights around the world. As part of its mission, Human Rights Watch pressures governments to prevent any form of torture or excessive use of force against people in their countries. The organization also urges governments not to transfer individuals to the custody of other countries where they are likely to be tortured. An international body affiliated with Human Rights Watch monitors compliance with human rights rules.

Next, we asked respondents to read a hypothetical scenario in which a monitoring body accuses the United States government of committing human rights violations. The identity of the monitoring party is matched to the initial description each respondent received. Respondents that viewed the Convention Against Torture (CAT) description were told that an international committee composed of either fifteen experts or fifteen countries identified the violation. Those that received the NGO description were told that the violation was identified by Human Rights Watch.

> Suppose that [**an international committee of fifteen independent experts/an international committee of fifteen countries/Human Rights**

Watch] issues a report describing how the United States violated the human rights of several individuals. The report alleges that the United States transferred several people to a country where US government officials knew they were likely to be tortured.

Finally, respondents were assigned to one of four conditions: a control condition (no additional information), a national security challenge, a bias challenge, or a sovereignty challenge. The language for these three shirking conditions is presented below.

National Security Challenge: In response to the report, US government officials argue that extenuating circumstances justified the violation because the government was responding to an extreme national security threat.

Bias Challenge: In response to the report, US government officials argue that the international body is clearly biased against the United States.

Sovereignty Challenge: In response to the report, US government officials argue that the international body has no right to challenge the actions of a sovereign nation like the United States.

After being randomly assigned to one of the conditions, respondents indicate their level of approval of the government's actions in the scenario on a 5-point scale. Possible responses include "strongly approve," "somewhat approve," "neither approve nor disapprove," "somewhat disapprove," or "strongly disapprove." This expression of approval is the primary outcome we use when analyzing results. In addition, we ask respondents their opinion on the likelihood that the report was accurate as well as the likelihood that the government will commit future human rights abuses. Finally, to assess whether respondents are willing to take costly action in response to the violation, we ask if they are willing to donate a portion

of their proceeds to a human rights advocacy group that promotes compliance with international rules prohibiting torture. In the remainder of the survey, respondents are asked to share basic demographic characteristics (e.g., age, race, income, etc.), indicate their political party identification, and answer a series of questions on their general disposition towards foreign policy.

4.2 Results

Government Approval

We first report the average treatment effect of each shirking strategy on respondents' approval of the government. Positive effects indicate a shirking strategy "works" from the government's perspective — i.e., it increases approval for the government compared to the counterfactual where the government provides no justification. A negative effect would suggest a shirking strategy backfired, leading respondents to lower their opinion of the government as a result of its justification.

Figure 1 displays the average treatment effect for each of the three shirking strategies across the three frames. The figure highlights several important findings. First, shirking is possible. Governments can mitigate the reputation cost of rule violations merely by adopting certain public postures towards a monitoring body. In three of the nine treatment conditions, there is evidence that the rhetorical defense offered by the US government increased the approval of respondents compared to the control condition. Specifically, the government experienced a statistically significant increase in approval when it employed a sovereignty challenge against an NGO, as well as a national security challenge against both an NGO and a monitoring body composed of IO member states. The substantive size of these effects is quite large. On a 5-point scale, the national security challenge in the NGO condition increases government approval by 0.79, or about two-thirds of the average difference between Democratic and Republican respondents.

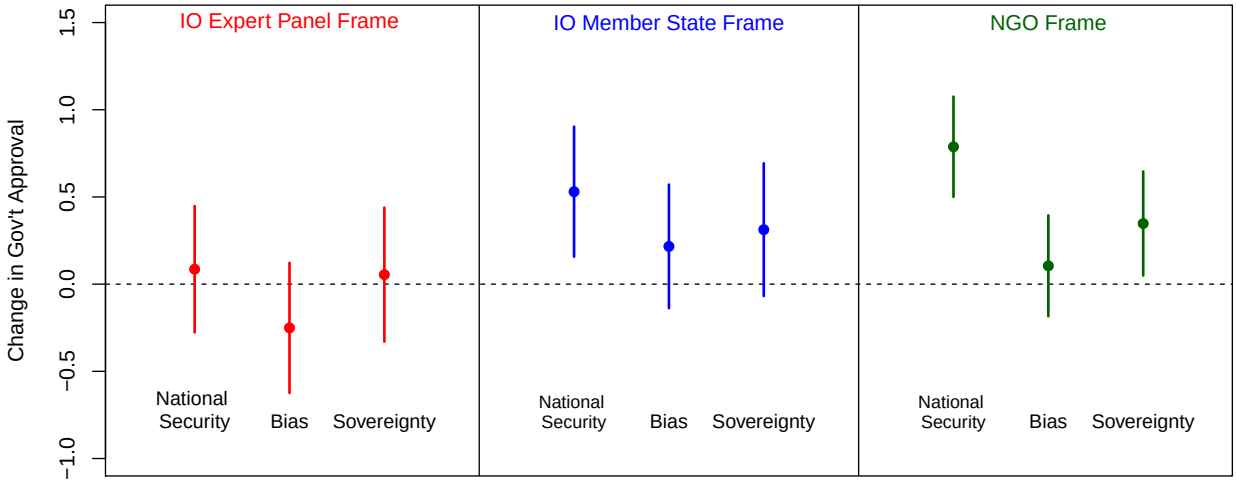

Figure 1: Effect of Shirking Strategies on Government Approval. The figure displays the average difference in government approval between each shirking strategy and the control group in the same frame, along with 95% confidence intervals.

Second, the results indicate that not all shirking strategies are equally effective. The invocation of national security concerns is the most powerful of the strategies examined here, prompting a statistically significant increase in approval in two of three experimental frames. Claims about state sovereignty are only effective in the NGO frame. The voluntary nature of international treaties likely reduces the effectiveness of an appeal to sovereignty, consistent with the observed insignificant effect in the two IO frames. The bias challenge has no significant effect on approval in any frame.

Third, there is some evidence that states' ability to shirk reputation cost is tied to the type of monitoring body that identifies a violation. When human rights commitments are monitored by a panel of independent experts, respondents are unswayed by shirking strategies. Monitoring behavior by member states or NGOs, on the other hand, seems to provide violating states with an opening to challenge allegations of non-compliance. This suggests an important advantage for institutions that delegate monitoring to independent experts: their ability to shape reputations is (relatively) immune from rhetorical attacks

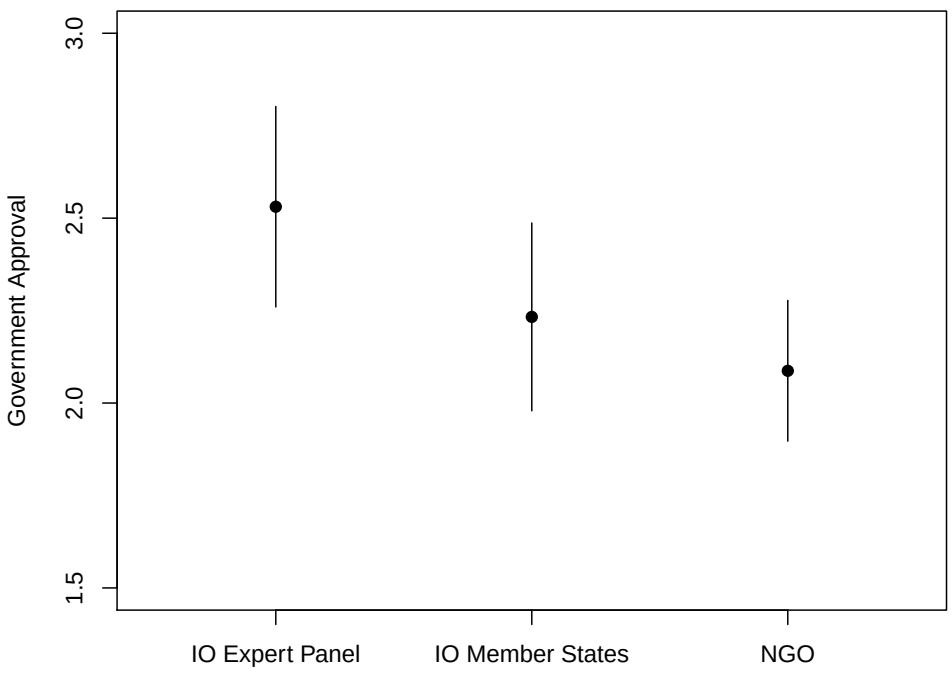

Figure 2: Government Approval across Monitoring Bodies. The figure displays the average baseline level of government approval when the government is accused of non-compliance by each type of monitoring body and the government offers no response (control condition).

launched by non-compliant states.

A deeper look at the results suggests a potential trade off to the power of independent monitoring bodies. Their insulation from from shirking strategies means that their effect on non-compliant states' reputations is immune to the public posture adopted by those states. It does not necessarily mean this type of monitoring body has a larger reputational effect to begin with. To compare the baseline reputational effect of each type of monitoring body, Figure 2 shows respondents' approval of the government in the control condition of each frame. In effect, this comparison gauges the power of each monitoring body when it sends an uncontested signal of non-compliance.

The figure reveals a surprising degree of variation in government approval across the three

frames. Government approval is highest when the government is accused of non-compliance by a panel of independent IO experts. Respondents assess the government in the harshest light when a violation is identified by an NGO. The difference in government approval in the IO expert and NGO frame is statistically significant ($p < 0.01$). Although we cannot speculate as to why respondents find an uncontested claim of non-compliance more persuasive from an NGO, this finding highlights a potential tradeoff between monitoring institutions. NGO monitoring may have a more powerful baseline effect on government approval, but it invites rhetorical challenges from governments that may eventually undermine its legitimacy. Monitoring by IO experts is comparatively weak but better able to withstand the shirking strategies states routinely employ.

Other Outcomes: Costly Action, Perceived Accuracy of Monitor

In addition to assessing the effect of shirking strategies on approval of the government, we investigated whether respondents would take costly action in response to an alleged violation. In particular, we asked each respondent whether they would be willing to donate a portion of their survey reimbursement to a human rights advocacy group that promotes compliance with international rules prohibiting torture. Respondents could indicate their willingness to donate 100% of their proceeds ($0.75), 80% ($0.60), 60% ($0.45), 40% ($0.30), 20% ($0.15), or none. Overall, more than 40% of respondents were willing to donate a positive amount to a human rights advocacy group. Figure 3 demonstrates the average effect of each shirking strategy on the amount of funds donated by respondents. Unlike in the government approval outcome, no shirking strategy generates a significant shift in donated funds compared to the control condition.

As above, we can also compare each monitoring body's ability to induce costly action from respondents when an accusation of non-compliance is uncontested. Figure 4 shows the average donation from respondents in the control condition for each monitoring body.

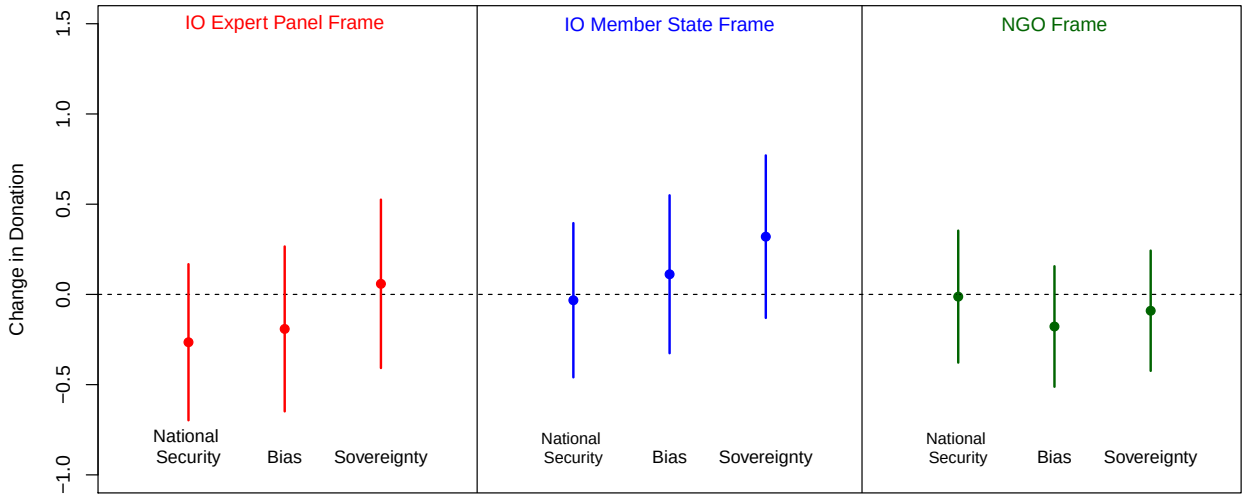

Figure 3: Effect of Shirking Strategies on Donated Funds. The figure displays the average difference in funds donated to a human rights advocacy between each shirking strategy and the control group in the same frame, along with 95% confidence intervals.

Unlike in the case of government approval, the panel of IO experts is the most consequential in generating costly action from respondents. On average, respondents in the control (no government response) treatment who were told that a panel of IO experts identified the human rights violation donated \$0.19 of their \$0.75 proceeds, compared to \$0.15 and \$0.16 in the IO member state and NGO frames. Although these differences are not statistically significant, the results provide some vindication for the IO expert panel as an effective monitoring body.

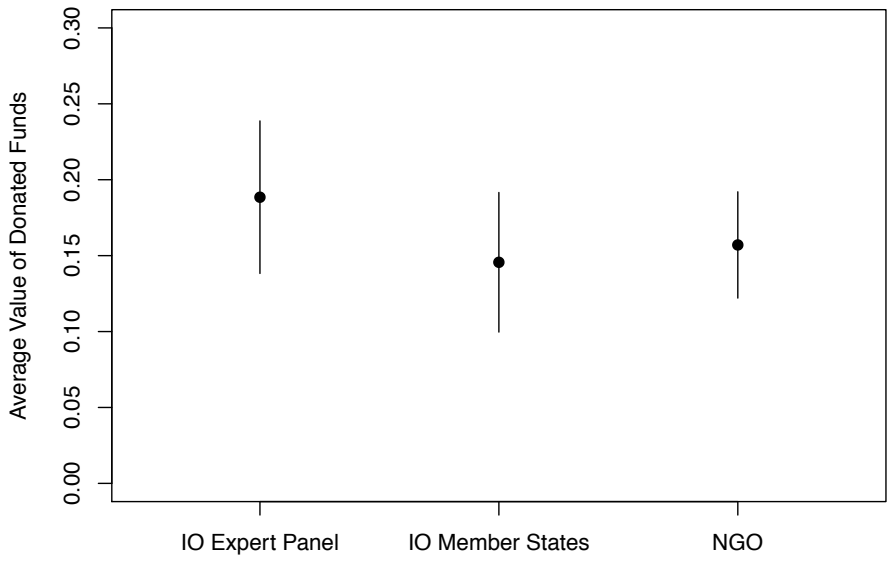

Figure 4: Donated Funds across Monitoring Bodies. The figure displays the average value of donated funds when the government is accused of non-compliance by each type of monitoring body and the government offers no response.

One possibility is that shirking works because respondents view monitoring as less accurate. To test this possible causal mechanism, we asked respondents in each condition to assess the accuracy of the monitoring report. Interestingly, we find little evidence to support this hypothesis. In general, respondents view the monitoring reports as similarly accurate in the control and shirking conditions, and this pattern transfers across all types of monitoring bodies (Figure 5). The one exception to the trend is when monitoring comes from an IO member state body and a government justifies its behavior as related to national security. In this scenario, respondents viewed the monitoring report as more accurate, perhaps because they interpreted the national security defense as a tacit acknowledgement of government action.

If shirking does not affect perceptions of accuracy, it may increase approval by expanding the context of reputation. Individual leaders and governments may have multiple reputations across different traits and policy issues. Our measure of governmental approval captures a

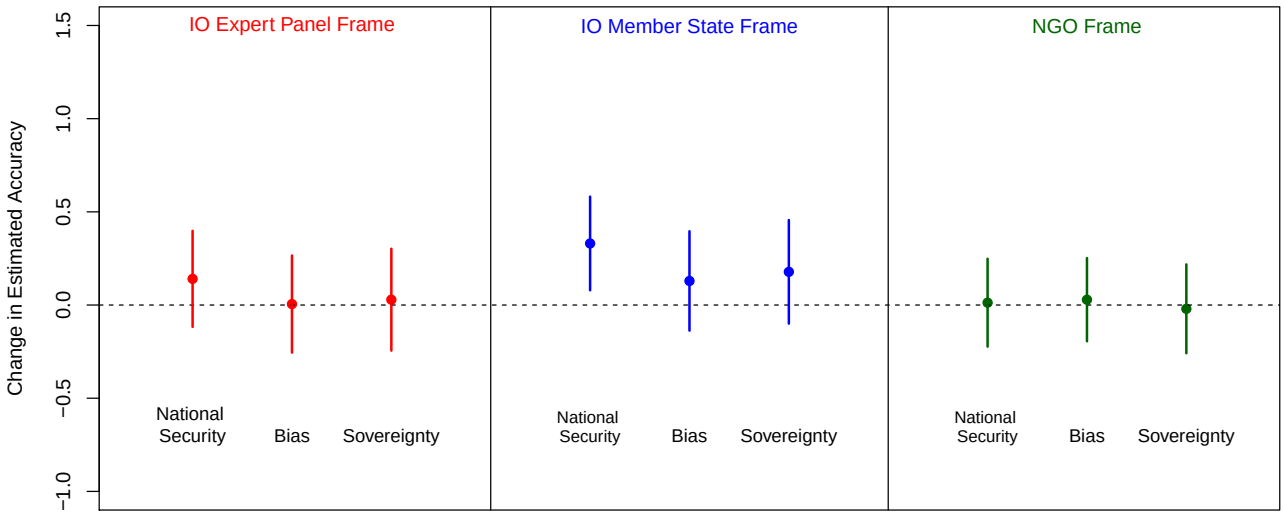

Figure 5: Effect of Shirking Strategies on Perceived Accuracy of Monitoring Body. The figure displays the average estimated accuracy of the monitoring report between each shirking strategy and the control group in the same frame, along with 95% confidence intervals.

broader aspect of government performance than simply the tendency of the government to comply with its international commitments. Perhaps shirking works by increasing the salience of other dimensions of reputation. There is some empirical support for this possibility. In the control condition, when the government provides no defense for its behavior, government approval is positively correlated with respondents' assessed accuracy of the monitoring report (0.098). This is consistent with respondents updating their opinion of the government based on the information they glean from the IO signal. When the government uses a shirking strategy, however, approval is weakly correlated with accuracy (-0.024). This may reflect the fact that respondents, primed by the government's rhetorical attacks, are prioritizing other dimensions when formulating their views about the government's performance.

Partisanship as a Moderating Variable

Finally, we examine whether partisanship moderates the effects of shirking strategies. In general, our results suggest that democrats and republicans behave similarly in response to a government's efforts to justify non-compliance.[12] In the IO expert condition (Figure 6), both democrats and republicans are resistant to shirking. In the IO member state condition and the NGO conditions (Figures 7 and 8), the results diverge slightly, particularly in the bias challenge, but differences are not statistically significant.

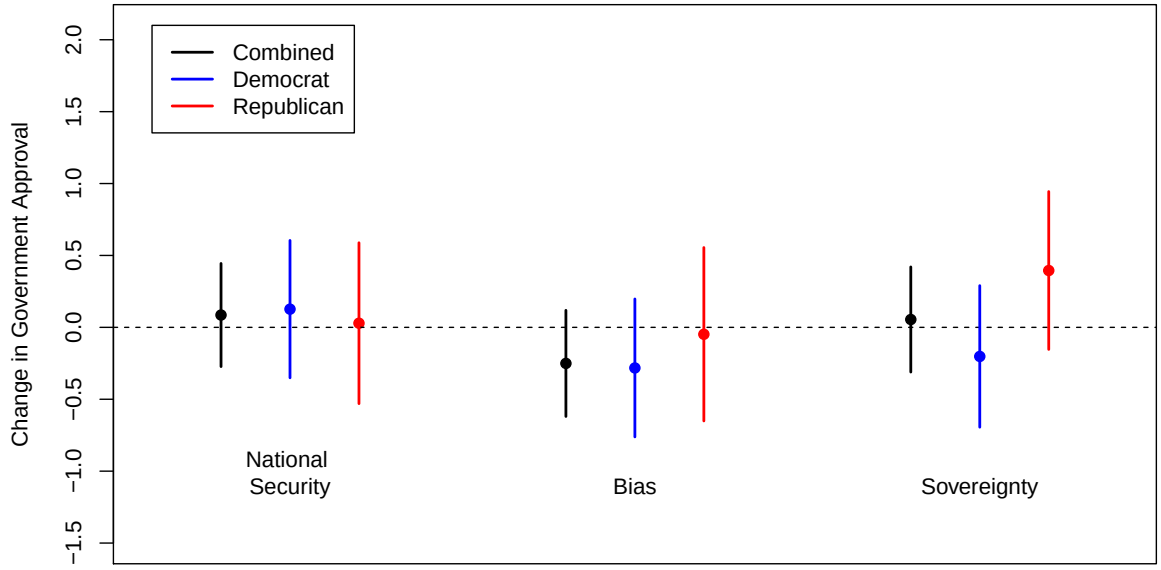

Figure 6: Effect of Shirking on Approval by Party ID, IO expert frame.

[12]We note that our conclusions our tentative due to power considerations.

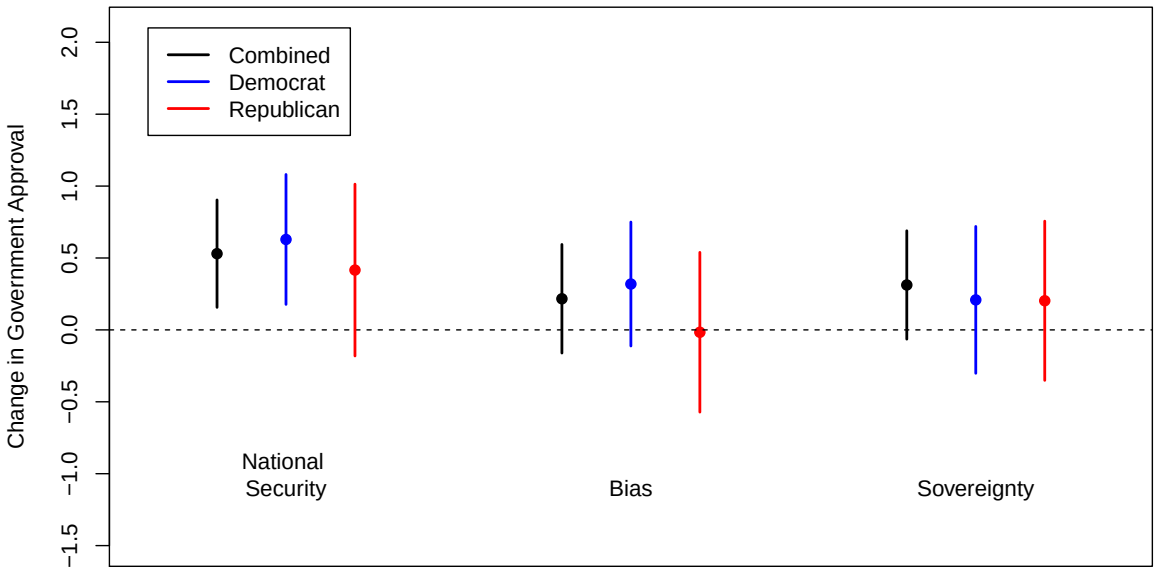

Figure 7: Effect of Shirking on Approval by Party ID, IO member state frame.

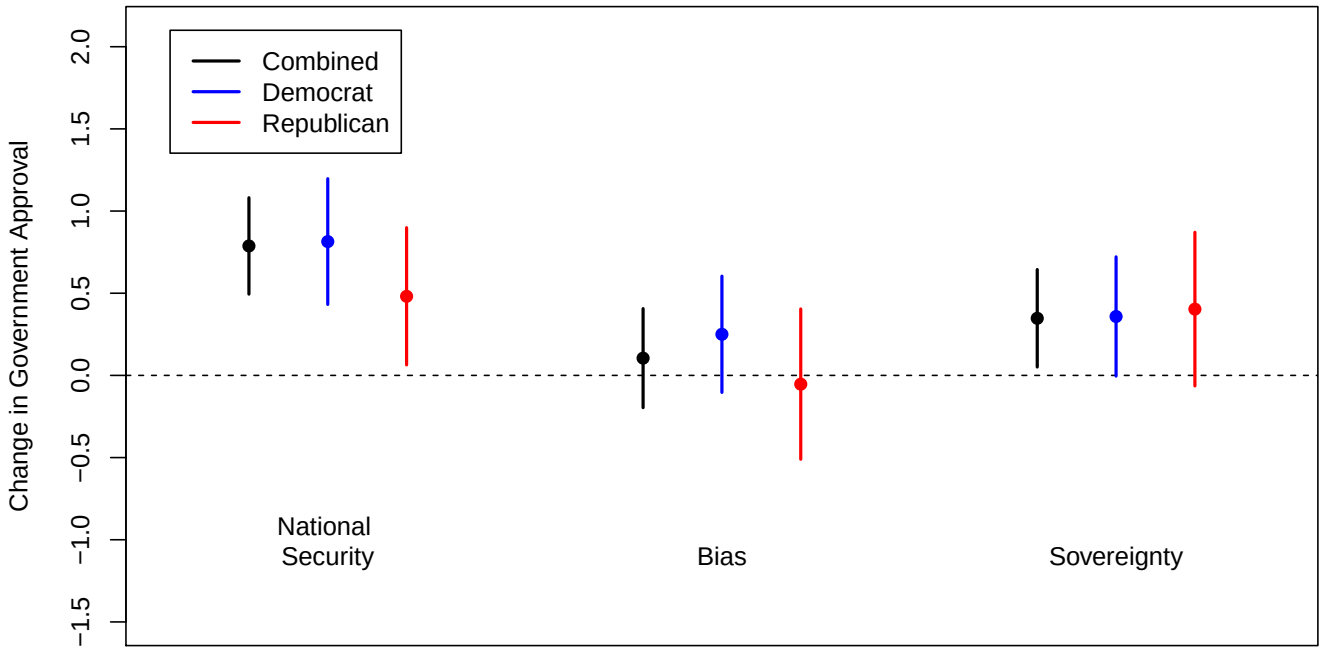

Figure 8: Effect of Shirking on Approval by Party ID, NGO frame.

5 Conclusion

As the complexities of globalization increasingly demand transnational solutions, the question of how to encourage compliance with international rules becomes ever more relevant. Most IOs lack enforcement capabilities, but if institutions can promote compliance via reputational mechanisms, the outlook for future cooperation looks more promising. Governments, however, spend significant resources shaping how they are perceived internationally and domestically. Even when an IO sends a clear signal about non-compliant behavior, a government may be able to shirk reputational costs by justifying its actions with reference to national security or sovereignty. As the last few years of public discourse have made clear, information can be easily contested, distorted, or manipulated to tell truths that serve

political purposes.

But even if government shirking is probable, IOs often have the ability to mitigate the effectiveness of this strategy. When IOs monitor state compliance with international commitments, and when monitoring reports come from independent experts rather than member states, IOs can prevent shirking. We demonstrate this result empirically through a survey experiment of shirking on the context of human rights violations. While human rights has higher normative significance than many other types of commitments, and thus could potentially be more resistant to shirking, we show that both national security challenges and sovereignty challenges are effective at undermining the reputational consequences of noncompliance. Importantly, however, this effect only occurs when monitoring comes from an IO member state body or an NGO. When an IO expert group monitors compliance, respondents do not find national security or sovereignty a compelling justification for human rights violations.

Our findings have significant implications for international relations theories that rely on reputation as a key causal mechanism. Negative information about state behavior undoubtedly causes reputational damage among domestic audiences, but the precise context for when such effects occur may be more limited that previously imagined. A leader usually has the opportunity to explain his actions to constituents; rhetorical justifications may often be sufficient to mute reputational effects. Under certain conditions, however, an IO can prevent shirking and impose reputation costs on uncooperative states. When IO monitoring comes from an independent expert body, respondents are less willing to accept a government's justification. Delegation to experts may limit the control of individual member states over outcomes, but such a cost may be worth it if states are truly committed to seeing policy improvements across states.

References

Axelrod, Robert. 1984. *The evolution of cooperation*. Vol. 5145. Basic Books (AZ).

Barnett, Michael N, & Finnemore, Martha. 1999. The politics, power, and pathologies of international organizations. *International organization*, **53**(4), 699–732.

Baum, Matthew A. 2004. Going Private: Public Opinion, Presidential Rhetoric, and the Domestic Politics of Audience Costs in U.S. Foreign Policy Crises. *Journal of Conflict Resolution*, **48**(5), 603–631.

Berinsky, Adam J, Huber, Gregory A, & Lenz, Gabriel S. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, **20**(3), 351–368.

Bradley, Curtis A, & Kelley, Judith G. 2008. The Concept of International Delegation. *Law and Contemporary Problems*, **71**(1), 1–36.

Bradner, Eric. 2014. Former Bush officials defend interrogation tactics. *CNN*.

Brutger, Ryan, & Kertzer, Josh. 2018. A Dispositional Theory of Reputation Costs. *International Organization*, **72**(3), 693–724.

Chaudoin, Stephen. 2014. Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions. *International Organization*, **68**(1), 235–256.

Clifford, Scott, Jewell, Ryan M, & Waggoner, Philip D. 2015. Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology? *Research & Politics*, **2**(4), 2053168015622072.

Committee Against Torture. 2006. *Conclusions and recommendations of the Committee*

against Torture: United States of America. adopted at the 36th Session of the Committee Against Torture. CAT/C/USA/CO/2, 25 July 2006.

Cooley, Alexander, & Snyder, Jack. 2015. *Ranking the World: Grading States as a Tool of Global Governance*. New York: Cambridge University Press.

Dafoe, Allan, Renshon, Jonathan, & Huth, Paul. 2014. Reputation and status as motives for war. *Annual Review of Political Science*, **17**, 371–393.

Dai, Xinyuan. 2007. *International institutions and national policies*. Cambridge University Press.

Davies, Graeme A.M., & Johns, Robert. 2013. Audience Costs among the British Public: The Impact of Escalation, Crisis Type, and Prime Ministerial Rhetoric. *International Studies Quarterly*, **57**(4), 725–737.

Downs, George W, & Jones, Michael A. 2002. Reputation, compliance, and international law. *The Journal of Legal Studies*, **31**(S1), S95–S114.

Fearon, James D. 1994. Domestic Political Audiences and the Escalation of International Disputes. *The American Political Science Review*, **88**(3), 577–592.

Franck, Thomas M. 1990. *The Power of Legitimacy Among Nations*. New York: Oxford University Press.

Gronke, Paul, Rejali, Darius, Drenguis, Dustin, & Hicks, James. 2010. U.S. Public Opinion on Torture, 2001-2009. *Political Science and Politics*, **43**(3), 437–444.

Guzman, Andrew T. 2008. *How International Law works: A Rational Choice Theory*. Oxford University Press.

Hafner-Burton, Emilie M., & Tsutsui, Kiyoteru. 2005. Human Rights in a Globalizing World: The Paradox of Empty Promises. *American Journal of Sociology*, **110**(5).

Hawkins, Darren G, Lake, David A, Nielson, Daniel L, & Tierney, Michael J. 2006. *Delegation and Agency in International Organizations*. Cambridge University Press.

Holdsworth, Nick. 2015. Kremlin, Russian Media Slam Allegations of State-Sponsored Doping. *hollywoodreporter.com*, November 10.

Huth, Paul K. 1997. Reputations and Deterrence: A Theoretical and Empirical Assessment. *Security Studies*, **7**(1), 72–99.

Kelley, Judith G. 2017. *Scorecard Diplomacy: Grading States to Influence their Reputation and Behavior*. New York: Cambridge University Press.

Kelley, Judith G, & Simmons, Beth A. 2015. Politics by Number: Indicators as Social Pressure in International Relations. *American Journal of Political Science*, **59**(1), 55–70.

Kelley, Judith G, & Simmons, Beth A. 2018. Introduction: Global Assessment Power in the Twenty-First Century. *Working Paper*.

Keohane, Robert O. 1984. *After Hegemony*. Cooperation and Discord in the World Political Economy. Princeton University Press.

Kertzer, Joshua D., & Brutger, Ryan. 2016. Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory. *American Journal of Political Science*, **60**(1), 234–249.

Koh, Harold Hongju. 1998. Is International Law Really State Law? *Harvard Law Review*, **111**(7), 1824.

Kono, Daniel Y. 2007. Making anarchy work: International legal institutions and trade cooperation. *Journal of Politics*, **69**(3), 746–759.

Levendusky, Matthew S., & Horowitz, Michael C. 2012. When Backing Down is the Right Decision: Partisanship, New Information, and Audience Costs. *Journal of Politics*, **74**(2), 323–338.

Martin, Lisa, & Simmons, Beth. 2012. International Organizations and Institutions. *Pages 326–351 of:* Carlsnaes, Walter, Risse, Thomas, & Simmons, Beth A (eds), *Handbook of International Relations*.

Mercer, Jonathan. 1996. *Reputation and international politics*. Cornell University Press.

Milgrom, Paul R., North, Douglass C., & Weingast, Barry R. 1990. The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics & Politics*, **2**(1), 1–23.

Morrow, James D. 2007. When do states follow the laws of war? *American Political Science Review*, **101**(3), 559–572.

Morse, Julia C. 2017. Blacklists, Market Enforcement, and the Global Regime to Combat Terrorist Financing. *Working Paper*, Feb.

Neumayer, Eric. 2005. Do International Human Rights Treaties Improve Respect for Human Rights? *Journal of Conflict Resolution*, **49**(6), 925–953.

Quigley, John. 2009. The United States' Withdrawal from International Court of Justice Jurisdiction in Consular Cases: Reasons and Consequences. *Duke Journal of Comparative and International Law*, **19**, 263–306.

Risse, Thomas, & Sikkink, Kathryn. 1999. The socialization of international human rights norms into domestic practices: introduction. *In: The Power of Human Rights International Norms and Domestic Change*.

Sanders, Rebecca. 2011. (Im) plausible legality: The rationalisation of human rights abuses in the American global war on terror. *The International Journal of Human Rights*, **15**(4), 605–626.

Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, Thomas. 1966. *Arms and Influence*. New Haven: Yale University Press.

Simmons, Beth A. 2000. International law and state behavior: Commitment and compliance in international monetary affairs. *American Political Science Review*, **94**(4), 819–835.

Simmons, Beth A. 2009. *Mobilizing for Human Rights*. International Law in Domestic Politics. New York: Cambridge University Press.

Slantchev, Branislav L. 2006. Politicians, the Media, and Domestic Audience Costs. *International Studies Quarterly*, **50**(2), 445–477.

Stinson, Scott. 2016. Red-Faced Russia Lashes Out; Conspiracy theories run amok despite overwhelming and damning evidence. *Vancouver Sun*, December 12.

Tempest, Matthew. 2004. Blunkett defiant over anti-terror laws. *The Guardian*, 25 Feb.

Tomz, Michael. 2007. *Reputation and International Cooperation: Sovereign Debt Across Three Centuries*. Princeton: Princeton University Press.

Trager, Robert F., & Vavreck, Lynn. 2011. The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party. *American Journal of Political Science*, **55**(3), 526–545.

Weeks, Jessica. 2008. Autocratic audience costs: Regime type and signaling resolve. *International Organization*, **62**(1), 35–64.

Appendix

Shirking and Signaling – MTurk Survey

Start of Block: Introduction

This is a survey about foreign affairs and US foreign policy. It will ask some questions about foreign affairs and will ask for your opinion on a few policy issues. It will not ask for any personal information, like your name or address, and all the information you provide will be used confidentially. Mechanical Turk Worker IDs will only be collected for the purposes of distributing compensation, will be removed from the dataset, and will not be shared with anyone outside of the research team. Your participation is completely voluntary and you can decide at any time not to answer a question or all questions. You can stop this survey at any time.

Do you agree to participate in the study?

- ○ Yes, I agree to participate
- ○ No, I don't wish to participate

End of Block: Introduction

[Respondents who agree to participate are randomly assigned one of the following three frames (IO Experts, IO Member States, NGO)]

Start of Block: Opening text – IO Experts

The United States is a member of the Convention Against Torture, an international treaty that seeks to promote justice and human rights around the world. Under this agreement, governments commit that they will not allow any form of torture or excessive use of force against people in their countries. They also agree not to transfer individuals to the custody of other countries where they are likely to be tortured. An international body affiliated with the Convention monitors compliance with human rights rules.

You will now read a hypothetical scenario regarding allegations that the United States Government committed human rights violations. In the text on the next page, we will describe the scenario and then ask you a series of questions.

Page Break

Suppose that an international committee of fifteen independent experts issues a report describing how the United States violated the human rights of several individuals. The report alleges that the United States transferred several people to a country where US government officials knew they were likely to be tortured.

End of Block: Opening text – IO Experts

Start of Block: Opening text – IO Member States

The United States is a member of the Convention Against Torture, an international treaty that seeks to promote justice and human rights around the world. Under this agreement, governments commit that they will not allow any form of torture or excessive use of force against people in their countries. They also agree not to transfer individuals to the custody of other countries where they are likely to be tortured. An international body affiliated with the Convention monitors compliance with human rights rules.

You will now read a hypothetical scenario regarding allegations that the United States Government committed human rights violations. In the text on the next page, we will describe the scenario and then ask you a series of questions.

Page Break

Suppose that an international committee of fifteen countries issues a report describing how the United States violated the human rights of several individuals. The report alleges that the United States transferred several people to a country where US government officials knew they were likely to be tortured.

End of Block: Opening text – IO Member States

Start of Block: Opening text - NGO

HRW Human Rights Watch is a non-profit, non-governmental organization that seeks to promote justice and human rights around the world. As part of its mission, Human Rights Watch pressures governments to prevent any form of torture or excessive use of force against people in their countries. The organization also urges governments not to transfer individuals to the custody of other countries where they are likely to be tortured. An international body affiliated with Human Rights Watch monitors compliance with human rights rules.

You will now read a hypothetical scenario regarding allegations that the United States Government committed human rights violations. In the text on the next page, we will describe the scenario and then ask you a series of questions.

Page Break

Suppose that Human Rights Watch issues a report describing how the United States violated the human rights of several individuals. The report alleges that the United States transferred several people to a country where US government officials knew they were likely to be tortured.

End of Block: Opening text - NGO

[Respondents in each frame are then randomly to one of the following shirking strategies (National Security, Bias, Sovereignty, Control (no additional text)):

National Security Challenge

In response to the report, US government officials argue that extenuating circumstances justified the violation because the government was responding to an extreme national security threat.

Bias Challenge

In response to the report, US government officials argue that the international body is clearly biased against the United States.

Sovereignty Challenge

In response to the report, US government officials argue that the international body has no right to challenge the actions of a sovereign nation like the United States.

Start of Block: Beliefs

B1 Do you approve or disapprove of the government's actions in this scenario?

- ○ Strongly approve
- ○ Somewhat approve
- ○ Neither approve or disapprove
- ○ Somewhat disapprove
- ○ Strongly disapprove

B2 Please tell us in a few words about why you approve or disapprove

_____

_____

_____

_____

_____

B3 In your opinion, how likely is it that the report was accurate?

- ○ Extremely likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Extremely unlikely

B4 If the report was accurate, what would concern you the most about the violation?

- ○ Threat to yourself or your family
- ○ Immorality of action
- ○ Damage to US reputation internationally
- ○ Likelihood that the government will violate other international commitments
- ○ Likelihood that the government will violate domestic commitments
- ○ Other (please specify)

B5 In your opinion, what is the likelihood that the government described in the scenario would commit future human rights abuses?

- ○ Extremely likely
- ○ Somewhat likely
- ○ Neither likely nor unlikely
- ○ Somewhat unlikely
- ○ Extremely unlikely

B6 The Committee Against Torture is a real committee that monitors compliance with international rules prohibiting the use of torture. In previous monitoring reports, the Committee has described problems in the United States with transferring individuals to countries where they are likely to be tortured.

Would you be willing to donate a portion of your proceeds from this survey to a human rights advocacy group that promotes compliance with international rules prohibiting torture?

- 100 percent ($0.75) (1)
- 80 percent ($0.60) (2)
- 60 percent ($0.45) (3)
- 40 percent ($0.30) (4)
- 20 percent ($0.15) (5)
- 0 percent ($0.00) (6)

End of Block: Beliefs

Start of Block: Manipulation Check

M1 In the scenario you just read, what kind of organization reported that the US government committed a human rights violation? (If you don't remember, please select your best guess.)

- A human rights organization
- A group of 15 independent experts
- A group of 15 countries

M2 In the scenario you just read, what explanation did the US government provide for the reported violation? (If you don't remember, please select your best guess.)

- ○ No explanation
- ○ The United States faced an extreme national security threat
- ○ The international body is clearly biased against the United States
- ○ The international body has no right to challenge the United States

End of Block: Manipulation Check

Start of Block: Foreign Policy Orientation

F1 We will now ask you some questions about how you believe the United States should act when it engages with other countries. There is no right or wrong answer - we are simply interested in your opinion.

F2 Please indicate your level of agreement or disagreement for each item:

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
The United States should always do what is in its own interest, even if our allies object (1)	○	○	○	○	○	○	○
The United States needs to cooperate more with the United Nations (2)	○	○	○	○	○	○	○
The United States should contribute forces to international peace-keeping efforts (3)	○	○	○	○	○	○	○
The United States should take all steps, including the use of force, to prevent aggression by any expansionist power (4)	○	○	○	○	○	○	○

FP3 Please indicate your level of agreement or disagreement for each item:

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
The United States must demonstrate its resolve so that others do not take advantage of it (1)	○	○	○	○	○	○	○
The use or threat of force sometimes creates more problems than it solves by creating hostility (2)	○	○	○	○	○	○	○
In deciding on its foreign policies, the U.S. should take into account the views of its major allies (3)	○	○	○	○	○	○	○
The United States should concentrate less on other countries and more on our own national problems. (4)	○	○	○	○	○	○	○

End of Block: Foreign Policy Orientation

Start of Block: PK Transition

Transition Only two sections left! You will now be asked a few questions about current events. You may or may not know the answers. If you don't know, please select your best guess.

End of Block: PK Transition

Start of Block: Political Knowledge Questions

PK1 Who is the current Vice President of the United States?

- ○ Hillary Clinton
- ○ Mike Pence
- ○ Joe Biden
- ○ Dick Cheney

PK2 Which of the following countries is not a member of NATO?

- ○ United States of America
- ○ Russia
- ○ United Kingdom
- ○ Canada

PK3 Which of the following countries is one of the United States' biggest trading partners?

- China
- Haiti
- North Korea
- Venezuela

PK4 What percentage of its annual budget does the US government spend on foreign aid?

- Zero
- About 1%
- About 20%
- About 50%

PK5 What does the G-20 stand for?

- A group of twenty African countries that share a common trade policy
- A group of the twenty largest militaries in the world
- A group of the twenty largest economies in the world

PK6 The EU refers to what international organization?

- The European Union
- The Election Unit of the United Nations
- The Economic Union

End of Block: Political Knowledge Questions

Start of Block: Additional Questions

Almost done! We just have a few more questions about your background.

A1 What is your age?

- Under 18
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65 - 74
- 75 - 84
- 85 or older

A2 What is your gender?

- Male
- Female
- Other

A3 What racial or ethnic group best describes you?

- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- Other

A4 What is the highest level of education you have completed?

- Less than high school
- High school graduate
- Some college
- 4 year degree
- Post-graduate degree

A6 Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what?

- Republican
- Democrat
- Independent
- Something else

A6B Do you consider yourself closer to the Democratic party or the Republican party?

- Closer to the Democratic Party
- Closer to the Republican Party
- Not close to either party

A7 Are you registered to vote?

- Yes
- No
- Don't know/do not disclose

End of Block: Additional Questions