Persuasion and Backlash from International Law in Global Swing States

Stephen Chaudoin and Taegyun Lim

September 30, 2025

Abstract

When one country or international organization makes accusations about violations of international law, the intended audience is often third-party states who might support punishing the offender. When do these accusations persuade publics in those countries and when do they trigger backlash? We show that reactions to accusations about international law violations are consistent with a theoretical model that allows for both types of responses – persuasion and backlash – depending on the audience member's prior beliefs and trust in the information source. We provide evidence from large survey experiments in four global swing states: India, South Africa, Turkey, and Indonesia. Swing states are where persuasion or backlash matter most, since allegations about international law could conceivably tilt their support toward the accuser or the accused. In our survey experiments, when the International Criminal Court makes accusations that Russia violated international law, this persuades certain subsets of the population. When the United States makes an identical accusation, this fails to persuade, and often backfires, because of the United States' lack of credibility as an accuser. We further show how accusations affect perceptions of the *accuser*, not just the accused. We show a dynamic feedback loop, where information sent today can increase or decrease views of the credibility of the information source. This can then magnify or mute the effect of future accusations. Accusations by the ICC improve respondents' views of the Court's credibility. Accusations by the United States further undermine its credibility.

1 Introduction

Accusations that a state has broken international law are a prominent rhetorical argument in international relations. To name just a few, the United States and the International Criminal Court have both accused Russia of committing war crimes in Ukraine. South Africa and a subsequent ruling from the International Court of Justice accused Israel of committing genocide in Gaza. The very first thing the Iranian foreign minister said in condemnation of U.S. airstrikes in 2025 was that they were "a grave and unprecedented violation of… international law."[1] Accusations that a state has violated international law constitute a message from a sender – such as another state or an international organization – to a receiver or audience – such as third-party states. The receiver then decides whether to change its policies or behavior toward the accused state. When a state or international organization (IO) makes accusations about violations of international law, it hopes to convince audiences in the receiving state to help punish the offender. We take direct aim at a fundamental question about these accusations: are they persuasive, and if so, for whom? When do accusations change third parties' beliefs about the target state's guilt or shift public support for punishing the accused? Additionally, how do accusations shape attitudes about the *accuser* as well as the accused?

Current scholarship finds a dichotomy between persuasion and backlash. Persuasion occurs when an accusation changes the audience's beliefs about the target's guilt and increases support for punishment. Backlash occurs when the audience shifts their beliefs or support in the opposite direction from that intended by the accuser. We describe a theoretical model that accommodates both persuasion and backlash. Audience members differ in their prior beliefs about the state of the world and their perceptions of the accuracy of an information source. A prominent feature of the model is that it gives clear predictions for both persuasion and backlash among different audience members. The effects of an accusation depend jointly on the audience member's prior beliefs about the state of the world and perceived credibility of the information source. Persuasion is most likely for an audience member who trusts the credibility of the information source and does not already have prior beliefs that agree

[1] "The Latest: US claims strikes on Iran's nuclear sites caused severe damage but full impact unclear," *AP News*, June 22, 2025. https://apnews.com/article/israel-palestinians-iran-war-latest-06-22-2025-7ab46578cb56feecc16f4e4940a46e0a

with the piece of information sent. Backlash is most likely when the audience distrusts the source but does not already staunchly disagree with the information.

The model also describes how accusations can change audience beliefs about the *accuser*, as well as the accused. Accusations can change their beliefs about the credibility of the information source itself, in addition to the target. Messages that reinforce the audience member's prior beliefs upgrade her views of the information source, and vice versa.

We show that both persuasion and backlash occur among a critical set of third-party audiences: citizens in "global swing states." By global swing states, we mean states that are not firmly aligned with the West or its adversaries. An accusation that Russia violated international law won't change many minds in Norway or Russia. Their support for or opposition to sanctions against Russia are firmly entrenched. However, publics in global swing states are critical audiences, because their support or opposition for punishing the target state is not as strongly pre-determined by their country's geopolitical alignment. Their citizens vary greatly in their prior beliefs about the target of the accusation and trust in information sources making the accusation, like United States officials or prominent international organizations. They are an important place to study persuasion and backlash, because an accusation could be pivotal for moving their citizens from opposing to supporting consequences for the accused, or vice versa. The swing states of the world could potentially go either way, so the marginal effect of an accusation may be greatest in swing states.

We test the predictions from the model using large survey experiments in Turkey, India, Indonesia, and South Africa (N = 6,742). The experiments describe a critical case: how accusations against Russia for alleged war crimes in Ukraine shape public opinion about Russian guilt and support for sanctions or aid to Ukraine. We included pre-treatment measures of respondents' prior beliefs about Russia and various measures of trust in the International Criminal Court (ICC) and United States as information sources. Respondents were then randomly assigned to receive a prompt attributing the accusations to either the ICC or the United States, with a control group receiving no such prompt.

We first examine whether accusations shift citizens' beliefs and policy preferences at both the aggregate and subgroup levels. Our analysis at the aggregate level shows that accusations by the United

States backfired. They produced outcomes opposite to their intended effects. Overall, U.S. accusations *reduced* respondents' beliefs that Russia committed war crimes and lowered their support for their government imposing measures such as sanctions on Russia. By comparison, accusations by the ICC had more muted impacts. ICC accusations generally had positive but often insignificant effects on respondents' beliefs about Russian guilt and their support for sanctions, as well as military and nonmilitary aid to Ukraine.

We then show how aggregate analyses alone obscure how different groups of citizens – depending on their prior beliefs and trust in information sources – update their views because of accusations. Once we account for heterogeneity in prior beliefs about Russian guilt and the credibility of the accuser, we find evidence of *both* persuasion and backlash. The patterns of persuasion and backlash are generally consistent with the predictions of the theoretical model.

Second, we show how accusations alter citizens' trust in information sources. The very act of making an accusation affected respondents' views of the accusers themselves. U.S. accusations diminished respondents' trust in the United States as a credible source of information. In contrast, ICC accusations increased trust in the ICC as a source of information. These effects were also moderated by respondents' prior beliefs, as predicted by the model. Those skeptical of Russian guilt were more likely to downgrade their perceptions of the accusers, and vice versa.

The broader implication of the first set of findings is that accusations from untrusted sources can do more harm than good. Accusations from the United States were more likely to backfire than persuade. While ICC accusations are only persuasive to a certain subset of the audience, they are nonetheless more persuasive than U.S. accusations, with significant differences between the two. Foreign governments like the United States would benefit from channeling their messages through IOs to be more persuasive and avoid backlash.

The United States' current commitment to cratering its credibility abroad will exacerbate this problem. One effect of plummeting U.S. soft power would be the erosion of its government's credibility in the eyes of many audiences abroad. The consequences of their weak credibility go beyond feelings toward the U.S or soft power. A lack of credibility affects whether the U.S. can rally foreign

support for sanctions and aid for allies - things which are tied directly to the hard power and material consequences of geopolitical conflicts.

The second set of findings has broader implications because they show a dynamic feedback loop. Accusations can shape future persuasion by altering trust in the source. Accusations that are well-received could convince the target to support the source's goals and increase their trust in the source, which will make the source even more persuasive the next time it seeks to sway opinions. Accusations that fall on deaf ears lower trust in the source, making their messaging even less effective the next time around. The perceived credibility of the accuser is itself endogenous, shaped itself by the signals sent by the information source. ICC accusations can potentially build a well of legitimacy, which may make it easier to persuade audiences to support its goals. Any successes towards its goals likely increases its credibility, which shapes future reactions, making future successes more likely. Conversely, the United States might erode its own credibility even further with poorly-received accusations, making success less likely, and in turn, decreasing its credibility the next time around. Here too, from the perspective of rallying foreign support, it would be better for the United States to avoid bluster when it lacks credibility.

Beyond our specific substantive context, our approach to modeling audience reactions unites many common arguments in experimental work about heterogeneous treatment effects under a common theoretical framework. Many existing arguments can be classified as arguments about priors about the state of the world, perceptions of sources, or some combination of the two. We show how measuring prior beliefs and trust in an information source can provide direct evidence of the mechanisms underlying arguments about persuasion and backlash. Our approach also directly matches its empirical designs with a theoretical model that yields clear predictions about the heterogeneous effects of messages. Recent work on the effects of IOs and international law has emphasized contestation. Our approach demonstrates a theoretical model and measurement strategy for making crisp predictions about how contestation will play out across and within important audiences.

Finally, our research highlights the importance of global swing states. The last decade has seen a resurgence of great power competition, pitting the United States and its allies against Russia and

China. The United States has increasingly abdicated its leadership role in the international order, choosing instead to engage in bilateral negotiations on many trade and security issues. States that are less strictly aligned with either bloc face pressure to choose sides, adding to the geopolitical importance of swing states.[2] Understanding the conditions under which messaging from leader states and IOs can persuade swing states will be critical to predicting the future directions of their foreign policies.

2 Information from IOs and Governments

Messaging from international organizations (IOs) and public diplomacy from governments is often designed to persuade. A sender (eg an IO or a state) transmits a piece of information to an audience (eg citizens or elites in another country). The sender hopes to change the audience's beliefs about the state of the world and the appropriate action they should take. For example, when an Indian citizen learns that the ICC has accused Russia of committing war crimes, the Indian citizen is the audience, the ICC is the sender, and Russia is the target. The citizen has prior beliefs about the true state of the world: whether Russia has broken international law. The citizen also has beliefs about the trustworthiness of the messenger: whether the source's information correctly matches the state of the world. The information she receives potentially changes her posterior beliefs about this state of the world and whether she should therefore support some action as a consequence, like sanctioning Russia.

Messages like "the target broke international law" are hoped to have the sender's desired effect on the audience because the message provides a comparison between the alleged behavior of the target and a legal standard, which the audience presumably wishes to be upheld. Foundational theories about the effects of international messaging emphasize the importance of reactions in third-party states, not just the target state itself. The "boomerang model" envisions this type of persuasion, where NGOs describe objectionable behavior in the target state. They hope that their information will persuade external audiences to exert pressure on the offending government, which is a critical step toward compelling repressive governments to change course.[3] As Murdie and Davis (2012) observes, "shaming

[2] Fontaine and McKinley (2025).
[3] Keck and Sikkink (1998).

often leads to heightened pressure on the repressive regime from third-party states, individuals, and intergovernmental organizations." This accusation is the first step in a longer dance of messaging and counter-messaging between accusers and the accused.[4]

Studies in international organization (IO) approval also emphasize the importance of third-party audiences' reactions. Chapman (2007) argues that IO authorization serves as a mechanism for information transmission, shaping audiences' beliefs about the likely outcomes of leaders' foreign policies. For example, UNSC authorization of a U.S. intervention persuades other countries that military action is not malicious adventurism. Such authorization is viewed as particularly trustworthy when it diverges from the preferences of powerful member states. This suggests that the persuasive power of IO signals varies with perceptions of institutional bias. Thompson (2006) and Thompson (2015) highlight strategic information transmission: IO involvement conveys signals about a coercing state's intentions and the likely consequences of its actions to both foreign leaders and their publics. This information shapes the level of international support the coercer receives from the public in third-party states. The argument helps explain the role of the United Nations Security Council in shaping international support for the use of force. Mikulaschek (2023) also suggests that the adoption of a policy by the European Union increases public support for the EU's policies. Unanimity in the EU increases the trustworthiness of this signal. Cohen and Powers (2024) finds weak effects of accusations from the "wronged country," because they are viewed as less credible. Chu (2025) brings a theory of social identity suggesting that in-group cues from IOs can strengthen foreign public support for humanitarian intervention. Recent advances in this area have focused on how IOs use new media communication tools to attempt to reach mass audiences[5] and how accused governments can avoid domestic political costs with image management.[6]

Research on public diplomacy – government-sponsored communication campaigns – shares a similar structure.[7] A sender transmits information to foreign citizens with the goal of moving their attitudes in a particular direction. These signals can take many forms, including public statements,

[4]Morse and Pratt (2025), Zvobgo (2019), Chow and Levin (2024).
[5]Carnegie, Clark, and Fan (2024)
[6]Morse and Pratt (2022).
[7]Goldsmith and Horiuchi (2009).

diplomatic visits, and endorsements. These campaigns are designed to shape international opinion by improving perceptions of the state's leaders, its people, and its core values, while also promoting support for specific foreign policy goals.[8]

Among all possible third-party states, we focus on a set we call "global swing states." We borrow the term swing state from its common usage in American politics because of the clear parallels. In American politics, swing states (like Pennsylvania or Georgia) are so-called because they can swing back and forth between choosing Republican and Democratic candidates in Presidential elections. This is in contrast with solidly blue or red states, which usually vote for the same party in Presidential elections (eg Alabama has not voted for a Democratic presidential candidate since 1976 and New York has not voted Republican since 1984).[9]

In international relations, swing states are those that are not strongly aligned with either of the two major blocs. Blocs in international relations are not as clearly well-delineated as they are in a two-party majoritarian system. But they are nonetheless useful heuristics.[10] Since the onset of the Cold War, there have loosely been two blocs: the west and its adversaries. The West includes the United States, Western Europe, and like-minded democratic allies such as Japan. The West's primary adversaries have shifted in identity and salience, but they have generally included the major autocratic countries, Russia and China.

Countries can differ in the degree to which they swing and this can of course vary across issue areas. But a handful of countries emerge as consistently inconsistent. Their common feature is their heterogeneity. The most commonly mentioned global swing states are: India, Brazil, Indonesia, Turkey, and South Africa. Each may lean towards supporting the West, but they do not follow lock-step.[11] With respect to Russia, these countries did not formally participate in the economic sanctions following its invasion of Ukraine in 2023.[12] Nor did they contribute to foreign aid to Ukraine, except for limited

[8] Goldsmith, Horiuchi, and Matush (2021), p.1342.

[9] To the best of our knowledge, Fontaine and Kliman (2013) and Kliman (2012) are two of the earliest usages of the term. We were surprised at the lack of emphasis on this concept. The term swing state does not appear in the *APSR*, *AJPS*, or *IO* except to refer to U.S. states. In *ISQ*, only Lee (2022) uses this term to refer to countries based on U.S. State Department sources.

[10] The term "third world" originally referred to countries neither aligned with the U.S. nor Soviet bloc.

[11] Fontaine and Kliman (2013), Fontaine and McKinley (2025)

[12] Syropoulos et al. (2024).

military and humanitarian assistance.[13] To be sure, there are other countries that do not follow lockstep with either the West or its adversaries. But we focus our attention on those states that have more economic and military might, which makes their decisions more consequential. To continue the analogy to U.S. electoral politics, we focus on states akin to Pennsylvania because they have more electoral college votes than Maine. As Fontaine and McKinley (2025) write: "None wishes to be forced into a strategic alignment with one great power alone… [Each] plays a dominant role in its region and takes actions with worldwide repercussions (4)."

Some work considers the effects of international law messaging on attitudes in global swing states. Suong, Desposato, and Gartzke (2024) survey respondents in Brazil (as well as Sweden, China, and Japan) and find that prompts about UN approval generally increase support for another country's use of force. Burcu Bayram, Keels, and Tokdemir (2024) study a survey experiment in Turkey (and the United States and Germany). They randomized aspects of an accusation about a foreign government's human rights practices, then measured how severely respondents thought that country should be punished. Respondents wanted allies punished less harshly than adversaries. Cope and Crabtree (2020) find that prompts about international law obligations regarding refugees backfire in Turkey, especially among incumbent party supporters. Chaudoin (2023) and Mikulaschek and Parizek (2025) describe the effects of the ICC and UN General Assembly resolutions on public opinion about international law. The former shows how the ICC shifted the content of media coverage of the war on drugs to more greatly emphasize human rights, but it also increased the degree to which contestation about the war received coverage. The latter show how a UNGA resolution condemning the Russian invasion of Ukraine decreased approval of Russian leadership in foreign countries where the media coverage of the UN was one-sided. In other countries, it had a more muted effect because of the uneven quantity and content of coverage across countries. Bassan-Nygate et al. (2024) find that international law treatment effects were similar in surveys conducted in seven countries, including India and Nigeria.

Among studies of public diplomacy, Mattingly et al. (2024) surveys numerous countries, including several swing states, comparing the effects of U.S. and Chinese propaganda. Mattingly and Sundquist

[13] Ukraine Support Tracker. https://www.ifw-kiel.de/topics/war-against-ukraine/ukraine-support-tracker/

(2023) shows how positive Chinese propaganda persuades Indian respondents, but negative attacks on the United States backfire. Morse and Pratt (2025) analyze a survey experiment answered by the U.S. public and a sample of 300 global elites, including some from swing states. They measure support for punishing an unnamed country accused of violating international law. Retorts by the accused generally decrease respondents' willingness to punish them, while IO rebuttals blunt some of these retorts.

2.1 Persuasion and backlash findings

A second key feature of existing evidence on the persuasiveness of information about international law is that their effects are often mixed. In addition to the persuasion and backlash examples in the preceding section, many other studies find one effect or the other, and sometimes both. Some studies find that IOs can persuade public opinion by providing credible information that aligns public beliefs with the content of the signal.[14]

However, IO signals can also trigger backlash, prompting individuals to reject the information or shift their beliefs in the opposite direction.[15] Negative predispositions toward IOs may lead some individuals to update their beliefs in opposition to IO messages, and publics may resist IO interventions when they directly target their own country.[16] Other studies highlight the importance of relational dynamics, suggesting that the effects of IO signaling depend on the relationship between the sender and the target—such as geopolitical alignment and perceptions of the sender as part of the in-group or out-group[17] These mixed findings suggest that audiences respond to IO signals in divergent ways.

Existing research suggests that public diplomacy can also shape foreign citizens' perceptions differently – producing persuasive effects in some contexts while provoking backlash in others. For example, diplomatic visits by foreign leaders have been shown to improve public perceptions of the visiting

[14]For example, information disseminated by IOs can increase public support for particular policies across a range of issue areas, including war (Grieco et al. 2011), military coalitions (Recchia and Chu 2021), migration policy (Mikulaschek 2023), and human rights (Anjum, Chilton, and Usman 2021).

[15]Cope and Crabtree (2020), Efrat and Yair (2023), Helfer and Showalter (2017), Voeten (2020).

[16]Bearce and Cook (2018), Chapman and Chaudoin (2020).

[17]Terman (2023), Pauselli (2023).

country's leadership and increase support for its security policies.[18] Similar positive effects have been documented for international organizations: visits by the UN Secretary-General, for instance, have been associated with improved human rights practices in host countries.[19] In contrast, other studies demonstrate that such efforts can backfire. Goldsmith and Horiuchi (2009) finds that diplomatic visits may provoke backlash, particularly under conditions of low source credibility or when perceived as strategic manipulation. Likewise, Rhee, Crabtree, and Horiuchi (2024) shows that public diplomacy can produce adverse reactions when foreign citizens suspect ulterior motives, due to psychological mechanisms such as insincerity aversion.

3 Theory

In this section, we describe a theoretical framework to explain how accusations can generate both persuasion and backlash among audiences. When individuals receive information about alleged misconduct, the effect on their posterior beliefs about the accusation depends jointly on their prior views and the perceived credibility of the information source. Our framework departs from conventional arguments that usually make predictions about the aggregate impact of accusations in a country. By formalizing belief updating, we show that aggregate approaches can obscure substantial heterogeneity in how different audience segments respond. We also show how this framework is versatile. It accommodates many disparate mechanisms for heterogeneous treatment effects that are posited in existing research. We are certainly not the first to argue that prior beliefs or perceptions of sources matter.[20] However, our framework shows how to make precise predictions about the moderating effects of these parameters and how measuring both is necessary to test many arguments.

Beyond belief updating about the target, the theory shows how accusations reshape perceptions of the information source itself. When accusations are perceived as credible, they enhance trust in the sender over time, creating a positive feedback loop. When perceived as biased, they erode trust and diminish the sender's future persuasive power.

[18] Goldsmith, Horiuchi, and Matush (2021), Wang et al. (2023).
[19] Choi et al. (2023).
[20] See for example Kertzer, Rathbun, and Rathbun (2020), Lupia and McCubbins (1998), and Grieco et al. (2011)

3.1 A theoretical framework for persuasion and backlash

We will continue with the example of an accusation about Russian war crimes for simplicity and to match the subsequent experimental setup. We assume there is a binary state of the world that is unknown to a citizen. Denote the state of the world as $S \in \{0, 1\}$, where $S = 1$ describes a situation where the accused is guilty. They have, in fact, committed war crimes. $S = 0$ denotes that they have not committed war crimes.

We call citizens the "audience." Each individual audience member, *i*, believes that the state of the world is drawn from a Bernoulli distribution, where the probability that $S = 1$ is $\pi_i \in [0, 1]$. Audience members therefore have heterogeneous prior beliefs about the probability that Russia is guilty, π_i.

The audience members all receive a common signal about the state of the world from a source. Let s_1 indicate that the source has sent a signal that $S = 1$, i.e. the source has said "Russia is guilty."[21] Audience members also have heterogeneous prior beliefs about the accuracy of the source: $\sigma_i = Pr(s_1|S = 1) = Pr(s_0|S = 0)$. In other words, σ_i denotes the audience member's prior beliefs that the source will send a signal that correctly matches the state of the world. We assume that each individual has prior beliefs about the accuracy of the information source. For individual i, her priors are that σ_i is distributed according to a Beta distribution with parameters α_i,β_i.[22]

We are interested in the *treatment effect* of a signal, s_1, on the audience member's posterior beliefs about two things: (1) the state of the world and (2) source accuracy. Applying Bayes rule, her posteriors about the state of the world are $Pr(S = 1|s_1) = \frac{\pi_i \alpha_i}{\pi_i \alpha_i + (1-\pi_i)\beta_i}$. Her posteriors about the accuracy of the source have a Beta distribution.[23]

[21] Note, this can incorporate "guilt" meaning "the accused did the act and it is illegal" and innocence as "they didn't do it" *or* "they did it, but it wasn't illegal.'' Our framework fits with either.

[22] Beta distributions are bounded between zero and one. They also have an intuitive link to prior beliefs about source quality. The expectation of source accuracy for an audience member is the proportion of signals from that source that correctly match the state of the world: $\mathbb{E}[\sigma_i] = \frac{\alpha_i}{\alpha_i + \beta_i}$. This is equivalent to an audience member who counts up the number of times the source has been correct in the past (α_i) and the number of times the source has been wrong (β_i), and then uses the proportion of correct past signals as her prior for beliefs about source accuracy.

[23] The fact that her posteriors about σ are distributed Beta follows from the Beta-Bernoulli conjugacy. The expectation of her posteriors about source accuracy are $\mathbb{E}[\sigma_i \mid s_1] = \Pr(S = 1 \mid s_1) \cdot \frac{\alpha_i + 1}{\alpha_i + \beta_i + 1} + \Pr(S = 0 \mid s_1) \cdot \frac{\alpha_i}{\alpha_i + \beta_i + 1}$. Proofs for all derivations are in the appendix.

We want the treatment effect to describe how these posteriors move relative to the audience member's priors. In other words, we want to think about the difference between her priors and her posteriors, not just her posteriors. This is a critical quantity of interest, because it describes how much the signal changes the audience member's beliefs. This also has a natural mapping to experimental and observational work about the effects of signals. We want to compare beliefs in a world where the audience member receives a signal, compared to a control condition where she does not receive a signal. In the latter case, without any signal, her posteriors are simply her priors.[24] In a between-subjects experimental design, researchers compare posteriors from a group that has been treated with some piece of information to a control group that has not received that information, and therefore retains their prior beliefs. In a within-subject design, researchers analyze the aggregate differences between pre-treatment (prior) beliefs and post-treatment (posterior) beliefs. In other words, our theoretical definition of "treatment effect" matches exactly the quantity of interest that is implied by nearly all experimental and observational data applications.

A key concept is how treatment effects differ, depending on the audience member's prior beliefs. We want to explicitly consider how the same piece of information can have different effects on different audience members. We use capital Greek letters to denote treatment effects. The treatment effect for the state of the world for audience member i is: $\Pi_i = \Pr(S = 1|s_1) - \pi_i$. The treatment effect for source accuracy is: $\Sigma_i = \mathbb{E}\left[\sigma_i|s_1\right] - \mathbb{E}\left[\sigma_i\right]$.

3.2 Updating about the state of the world

The relationship between priors and treatment effect predictions are easiest to see visually.[25] Figure 1 shows the predicted treatment effects for posteriors about the state of the world (Π_i). The horizontal

[24] It is worth noting that this description of a treatment effect does not capture updating in the absence of a signal. In other words, the audience member does not say "I haven't heard the source say anything about the state of the world, and the absence of that information is itself informative about the state of the world." Our exclusion of this directly matches experimental settings, where the researcher can strictly control the absence of a signal and the respondent does not know the information she could have received but did not receive. It is also likely to be a good first approximation of what happens outside of an experimental laboratory. News consumers have their preferred outlets and the consume the news that source provides to them each day. But they are not often thinking about all of the articles the source could have chosen to write but did not write.

[25] The expression for this quantity is: $\Pi_i = \frac{\pi_i \alpha_i}{\pi_i \alpha_i + (1-\pi_i)\beta_i} - \pi_i$. The derivation is in the appendix.

axis shows respondent i's prior beliefs about source accuracy. The vertical axis show her prior beliefs about the state of the world. For each cell in the plot, we calculate Π_i and the heatmap shows the magnitude and direction of the treatment effect.[26]

Blue cells on the right hand side indicate persuasion, where $\Pi_i > 0$. The audience member's posteriors have moved in the sender's intended direction. The bottom right quadrant is where persuasion is most powerful. Individuals in the quadrant didn't think Russia was guilty and they trust the accuracy of signal. They therefore show the greatest positive movement from their priors to their posteriors. The top right quadrant shows a ceiling effect, where the sender is "preaching to the choir." These individuals trust the signal, but they already thought Russia was guilty, so their posteriors are only a small increase over their priors.

Red regions indicate backlash, where $\Pi_i < 0$. These individuals think the signal is worse than uninformative. They think the signal should be interpreted in the opposite direction from the sender's intended effect. In the top left, these individuals thought Russia was guilty but distrust the signal, so they are the most "dissuaded" to believe the sender. In the bottom left, the sender's signal has "fallen on deaf ears." They distrust the signal, but there is a floor effect because they already didn't think Russia was guilty.

[26]Note, too, that this is equivalent to the treatment effect if we simulated many respondents in each cell, randomly assigned them to treatment or control, and then regressed their posterior beliefs on an indicator for treatment assignment. It other words, the heatmap also shows the predicted regression coefficient for the regressions used in most empirical studies.

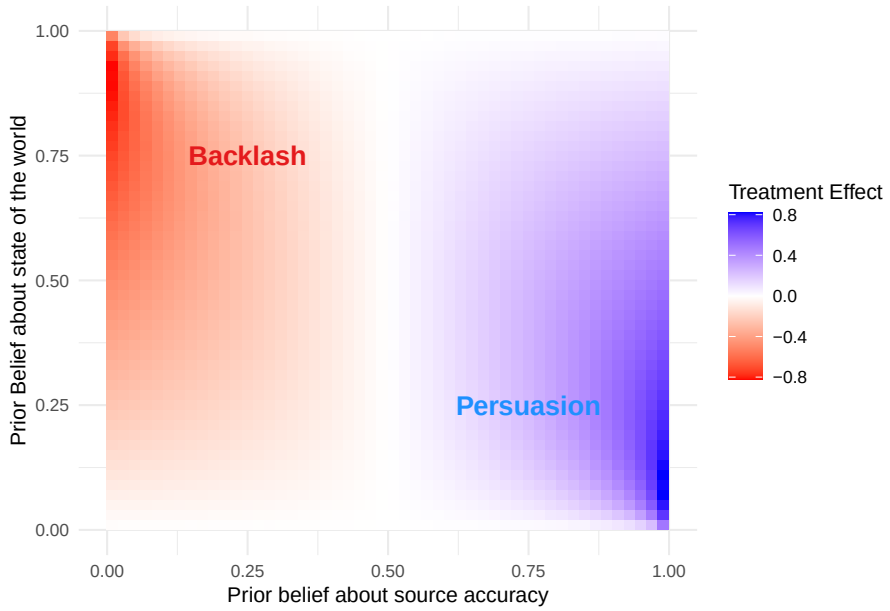


Figure 1: Predicted treatment effects on beliefs about the state of the world.

If audience members update their beliefs in a manner consistent with the model, then the magnitude and direction of belief change depend on both their prior beliefs and their trust of information sources. Some will be persuaded and others will move in the opposite direction to the signal.

H1 (Treatment effects on posteriors about the state of the world):

H1a (Persuasion): When individuals exhibit high trust in the source and have priors that are not already strongly aligned with the sender's intended direction, their posterior will move in the intended direction, to the greatest degree.

H1b (Backlash): When individuals exhibit low trust in the source and have priors that are already more aligned with the sender's intended direction, their posteriors will move in the unintended direction, to the greatest degree.

The difficulty of testing aggregate hypotheses, like "the signal persuades," without measurements of priors and likelihoods is apparent in Figure 1. If the sample included people evenly spread across all four quadrants, and a researcher regressed posterior beliefs on whether the individual got the signal, the coefficient would equal zero, *despite these individuals responding exactly as predicted in the model*! We might conclude that the signal was ignored, even if every individual was a "complier" with the

treatment and reacted in the exact way predicted by the model.

Figure 1 also shows why prior beliefs and perceptions of accuracy matter *jointly* in predicting treatment effects and for assessing heterogeneous treatment effects. Varying prior beliefs generally has a non-monotonic effect on the magnitudes of predicted treatment effects. For a given perception of signal accuracy, the treatment effects increase and then decrease when we move from the bottom of the figure to the top. And the gradient of the treatment effect as we vary priors also depends on whether the source is perceived as accurate or inaccurate. On the left hand side (signal is inaccurate), moving from bottom to top makes treatment effects more negative, and then less negative as priors converge to the limit. On the right hand side (signal is accurate), the opposite is true.

Going from left to right in the figure, increasing the accuracy of the signal unambiguously raises the treatment effect. But the magnitude of this change depends greatly on prior beliefs. When the receiver believes the target to be guilty (upper half), increasing accuracy changes the treatment effect from strongly negative to weakly positive. When the receiver believes the target to be innocent (lower half), increasing accuracy changes the treatment from weakly negative to strongly positive.[27]

Some existing studies consider priors or perceptions of accuracy in isolation. For example, Chaudoin (2014) finds that treatments about a possible WTO dispute have the largest effect for those who are neither strongly supporting of or opposed to free trade, *ex ante*.[28] Arguments about floor and ceiling effects are also arguments about prior beliefs.[29] The top right of Figure 1 is the canonical "ceiling effect", where treatments matter less because the receivers priors are already aligned with the message. The bottom left is the canonical "floor effect." Other existing studies consider moderators tied to the perceived accuracy of a source, again in isolation. For example, Anjum, Chilton, and Usman (2021) find larger effects of United Nations endorsements on Pakistani respondents who express trust in the United Nations.[30]

As we describe more extensively below, our empirical approach will aggressively measure respon-

[27] For a comparison between our model and motivated reasoning models, see the appendix.

[28] Spilker, Nguyen, and Bernauer (2020) also find that the effect of new information about a trade agreement is less impactful for those with stronger priors.

[29] E.g. Búzás and Bassan-Nygate (2024) and Cope (2023).

[30] For a recent survey, see Morrison (2024). Other examples include Mikulaschek (2023) and Bearce and Cook (2018).

dent priors and their perceptions of information sources. This enables us to examine reactions across the prior and accuracy space, and compare treatment effects with a concrete prediction for each type of respondent.

Figure 1 also makes it apparent why some moderators used in studies of heterogeneous treatment effects have ambiguous implications. Consider whether someone holds cooperative internationalist (CI) foreign policy attitudes. Scoring higher on a CI scale might make a respondent more trusting of an IO's information, but it can also blunt treatment effects if it means that the respondent already believes what the IO is telling them. Which effect dominates would be difficult to know *ex ante*. Their net effect is theoretically unclear.[31]

3.3 Updating about the source

The signal sent affects posterior beliefs about the accused, but it also can affect beliefs about the accuracy of the *source* itself.[32] The impact of information extends beyond its immediate persuasive effects, influencing long-term perceptions of the source itself. When the audience perceives a source as providing trustworthy information, they are more likely to increase their support for and trust in that source. Conversely, when information is initially perceived as biased, not only is the content of the message rejected, but this leads to a further erosion of trust . This loss of credibility can have significant downstream effects, diminishing the source's ability to shape public opinion or mobilize support in future interventions. Thus, the relationship between public perception and the credibility of the messenger creates a reinforcing dynamic: trustworthy sources build support over time, while untrustworthy sources face declining influence with continued engagement.

The model above also generates predictions about how the audience will update their beliefs about the accuracy of the source.[33] Figure 2 shows these predicted treatment effects for Σ_i.

The treatment effects are monotonically related to priors about the state of the world. For a respon-

[31] We demonstrate this ambiguity empirically in the appendix.

[32] Cheng and Hsiaw (2022), Gentzkow, Wong, and Zhang (Forthcoming).

[33] The expression for this treatment effect is $\Sigma_i = \frac{\alpha_i}{\alpha_i+\beta_i+1} \cdot \left[1 + \frac{\pi_i}{\pi_i\alpha_i+(1-\pi_i)\beta_i}\right] - \frac{\alpha_i}{\alpha_i+\beta_i}$. The appendix shows the derivation.

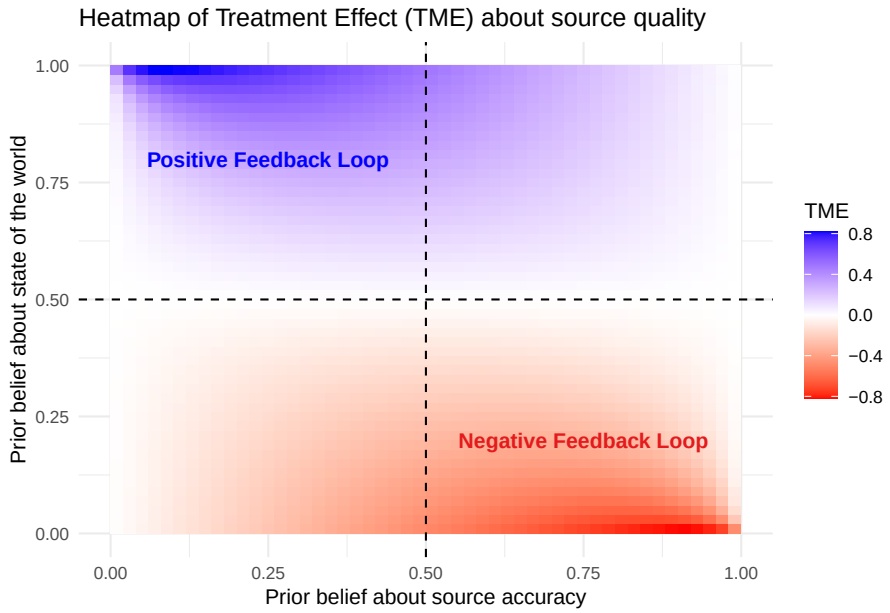


Figure 2: Predicted treatment effects on beliefs about the accuracy of the source

dent that has a particular prior about source accuracy, treatment effects are monotonically increasing in the prior beliefs that "guilty" is the state of the world. The intuition is that, if an audience member starts more convinced about the state of the world, and they receive a signal that comports with that prior belief, then they update favorably about the quality of the source. "If the source tells me what I think is already true, then I trust the source more."

The contours of the treatment effects are different from the predicted treatment effects about the state of the world. In the bottom right region, the audience member says "I thought you were an accurate source, but then you told me something that really contradicts my priors, so I lower my beliefs about your source quality." In the upper left, the audience member says "I thought you were a terrible source, but then you said something that matched my priors, so I upgraded my beliefs about you as a source." Hypothesis 2 describes these key features of the predicted treatment effects that we analyze empirically below.

H2 (Conditioning effect of priors about the state of the world): The signal sent by an information source can increase or decrease the audience's posterior beliefs about the accuracy of the source. The effect of a signal "guilty" on beliefs about the accuracy of the source are increasing in the individual's prior

beliefs that "guilty" is the state of the world.

The model shows the possibility that IOs or diplomacy can create a "positive feedback loop," if the public has at least some level of trust in the information source. Over time, sending information enhances the source's ability to persuade and shape public opinion, even if short-term effects are more limited. An information source can become more credible with consistent engagement that bolsters their legitimacy and long-term persuasiveness.

Some existing research examines how persuasive appeals influence perceptions of the message source. In the context of IOs, such messaging can shape public views of IO legitimacy.[34] In particular, different institutional characteristics – encompassing both procedural and performance-based qualities – can broadly shape public perceptions of the legitimacy of IOs. Relatedly, Brutger and Strezhnev (2022) and Chung (2025) show that information about disputes involving a respondent's country can erode public attitudes toward the IO associated with the dispute.

4 Experimental Design

4.1 Background and sample

We chose accusations against Russian war crimes for the context of our survey because it represents a watershed event in which global swing states have played a critical role. In March 2023, the International Criminal Court (ICC) issued arrest warrants for Russian President Vladimir Putin, alleging responsibility for war crimes committed during Russia's invasion of Ukraine in 2022. These accusations included the unlawful deportation of children from occupied territories in Ukraine to Russia, which the ICC classified as a violation of international law. Foreign governments, especially the United States, also condemned Russia for numerous instances of war crimes in Ukraine, ranging from indiscriminate attacks on civilian infrastructure to documented atrocities such as those committed in Bucha and Mariupol.

Reactions of swing states are especially important given the critical role these countries play in

[34] Dellmuth and Tallberg (2021), Ecker-Ehrhardt, Dellmuth, and Tallberg (2024), Ghassim (2024).

forming coalitions to implement punitive measures against Russia. Additional condemnation from France or support from Iran is largely irrelevant for Russia. Russia received condemnation or support from those countries before ICC accusations and continued to do so afterwards. Condemnation from a swing state like India, however, has much larger implications. For example, the effectiveness of economic sanctions on Russia largely depends on these states' willingness to enforce trade restrictions, limit access to financial systems, or reduce dependency on Russian energy exports.[35] Without the participation of swing states, sanctions are more easily circumvented, weakening their impact on Russia's economy and ability to sustain the war effort. Similarly, in other foreign policy areas—ranging from military and non-military aid to pure diplomacy—swing states' cooperation is crucial in exerting credible pressure on Russia.

We conducted survey experiments in four global swing states—Turkey (N = 1,664), India (N = 1,704), Indonesia (N = 1,672), and South Africa (N = 1,702)—in collaboration with TGM Research in October 2024. These countries do not consistently align with either Western powers or their primary geopolitical rivals, and they have not joined the sanctioning coalitions targeting Russia. At the same time, unlike states such as China or Belarus, they are less overt in efforts to undermine sanctions. Their cooperation is critical for enforcing costly international measures. For example, cooperation of these countries is important for blocking the rerouting of export-controlled items. Examining public opinion in these global swing states is therefore essential for building broad coalitions to enforce costly measures against violations of international law.

Figure 3 uses information from the 2021 World Gallup Poll surveys to show where these countries generally lie in their attitudes towards Russia and the United States.[36] The vertical axis shows the mean number of respondents indicating that they disapprove of the leadership of Russia. There is not a specific question about Russian war crimes, but approval of the leadership is likely correlated with views on the war in Ukraine. The horizontal axis shows approval of U.S. leadership. This, too, is

[35] Davis and Lim (2025). Since our survey, these states have taken on even greater importance with respect to sanctions on Russia. The U.S. has threatened India and Brazil with huge tariffs in an attempt to curb their oil imports from Russia. See "Putting maximum pressure on Russia requires secondary sanctions on oil" Washington Post, August 2, 2025, for example.

[36] The plot shows the top 25 countries by population in 2021.

likely correlated with views on U.S. credibility. The countries we chose are good examples of swing states because they have relatively moderate opinions towards Russia and the United States. They are not on the extremes of either distribution.

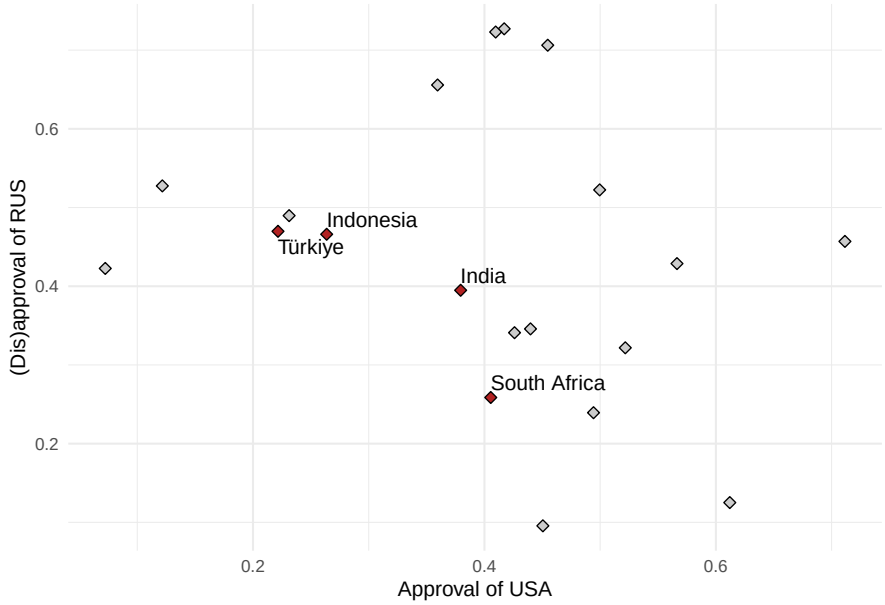


Figure 3: . Countries placed according to Gallup responses.

These countries have also typified the "non-aligned" stance of countries that neither actively support nor actively oppose Russia. Their overall stances have been neutral and non-committal. Table 1 summarizes each country's stance on some of the key issues surrounding Ukraine. Turkey is the only country to have provided direct military aid to Ukraine, though India has also given humanitarian assistance.[37] None of the four countries actively participate in the sanctions regime against Russia, though Indonesia stopped some arms imports from Russia and replaced them with French suppliers.[38] Each of the four countries either voted in favor of or abstained from the 2022 UN General Assembly resolutions condemning the latest Russian invasion of Ukraine. The countries also took tepid, generally non-committal stances on the ICC's arrest warrant for Putin. South Africa, especially, has tread carefully about the arrest warrants, since their ICC membership legally obliges them to arrest Putin should he visit the country.

[37] Source: IFW-Kiel Ukraine Support Tracker Data.
[38] Chivvis, Noor, and Geaghan-Breiner (2023).

Country	Aid to Ukraine	Sanctions Regime Participation	UNGA 2022 Resolutions	ICC Arrest Warrant
India	Non-military only	No	Abstained	Non-member, no public stance
Indonesia	None	No	Voted in favor	Non-member, no public stance
South Africa	None	No	Abstained	Member, mixed/critical public stance
Turkey	Some military (e.g. drones) and non-military	No	Voted in favor	Non-member, no public stance

Table 1: Positions of selected countries on Ukraine-related issues.

4.2 Pre-Treatment Measures

Pre-treatment, we measured respondents *prior beliefs* about whether Russia had violated international law. Our survey item read "Countries sometimes violate international laws of war that restrict attacking civilians and other acts. In your opinion, what is the percent chance that the countries below have violated international laws of war over the last 5 years?" Respondents chose from a sliding scale from 0-100. They answered for Russia, the United States and China.[39]

We also included three items that measure respondents' perceptions of the accuracy of a particular source of information. The first item asked "There are many sources of information about international affairs. Some sources of information are trustworthy and others are not. On a scale of 1-100, with zero being the least trustworthy and 100 being the most trustworthy, where would you place the following sources of information?" Respondents answered for the United States Government, the ICC, and the media.[40] The second item read "[Countries/international organizations] criticize each other. Sometimes they are telling the truth and other times they have another motive. In your opinion, what is the percent chance that these [countries/international organizations] are telling the truth

[39]We randomized the country order. Including the United States and China helps make this item not solely focused on Russia. We used the "percent chance" language since it has been used in previous studies conducted internationally that were focused specifically on measuring probabilities (Delavande (2014)).

[40]Again, we randomize the sub-items for this and all subsequent questions where applicable, and we included the media to avoid focusing solely on the two actors of interest.

when they criticize another country?". Respondents again chose on a scale from 0-100. The list of countries included the United States, China, and France. The list of IOs included the ICC, the WHO, and the EU. Third, we used a simple feeling thermometer for the United States and the ICC.[41]

In practice, for both sources of information (the ICC and the United States), all three measures of prior beliefs about the source — trustworthy, telling truth, and feeling thermometer—are strongly correlated. For the ICC, pairwise correlations range from 0.67 to 0.76, and for the United States, they range from 0.71 to 0.74. Given this high degree of internal consistency, we construct a single index for each source by taking the simple average of the three measures.

4.3 Treatment and Outcome Measures

Respondents assigned to the control group read the following sentence - *"As you may or may not know, Russia invaded Ukraine in 2022."* Respondents assigned to the U.S. or ICC treatment groups read the control group sentence, followed by an additional declaratory statement: *"The [United States/International Criminal Court] has accused Russian leaders of committing war crimes during the invasion."* We chose this treatment design because it is simple and minimal: the only new information it conveys is that a particular source has made an accusation.[42]

Many states condemned Russia. We selected the United States as the individual state treatment because it is widely regarded as one of the most influential countries in the world, both in material power or soft power. Thus, among the states that condemned Russia, accusations from the United States should be the most likely to persuade publics in global swing states.

Among international organizations that condemned Russia, we selected the ICC for the IO treatment because it is an independent legal institution, clearly distinguishable from the signaling of individual states. Unlike other high-profile IOs such as the UN or the EU—whose statements often reflect the collective positions of member states—the ICC operates through independent legal bodies,

[41]The item read "We'd like to get your feelings toward certain countries and international organizations on a 'feeling thermometer.' A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the country or organization at 50 degrees if you don't feel particularly positively or negatively toward them. How do you feel about following countries or international organizations?". The list also included Russia and Israel.

[42]Details on the manipulation checks are provided in the Appendix.

including a judiciary and an Office of the Prosecutor, which are not strongly associated with any single national interest. This makes the ICC a useful contrast for isolating the effects of signaling from a neutral, norm-enforcing IO versus an individual state actor.

We included three types of outcome measures: the respondents' posterior beliefs about Russian guilt, their preferences over policies toward Russia their country could adopt, and their beliefs about information sources themselves.[43] For posterior beliefs about Russian guilt, we asked "How likely is it that Russian leaders have committed war crimes in Ukraine?" and respondents used a 100-point scale.

For policy responses the respondent's government could adopt, we asked three agree/disagree questions about whether the respondent's government should: (1) "impose sanctions on the Russian government, companies, and individuals?", (2) "provide non-military aid to Ukraine?" and (3) "provide military aid to Ukraine?". Respondents chose from a five-point scale (Strongly agree/disagree, somewhat agree/disagree, neither agree nor disagree).

The overall level of support for the policy responses in each country was consistent with our characterization of them as swing states. If we assign numerical values to the 5-point agreement scale, 1-5, the mean of the responses across all four countries was 3.3 for non-military aid, 2.9 for military aid, and 3.0 for sanctions.[These numbers are calculated from control group respondents, since these measures were post-treatment. See appendix for country breakdowns.] Indian respondents had the strongest support for non-military and military aid (means of 3.4 and 3.2 respectively). South Africa had the strongest support for sanctions (mean of 3.2). Indonesia had the lowest means for all three policies (3.3, 2.6, and 2.8).

To assess post-treatment beliefs about information sources, we measured two outcomes: trust in the information source and perceived legitimacy of the source. For trust, respondents were asked: "Some sources of information are biased and others are not. On a scale from 0 to 100, with 100 being the most biased, where would you place the following sources of information?" Participants rated the ICC, the U.S. government, and the media.

[43]We randomized the order of these items across respondents, following advice from Chaudoin, Gaines, and Livny (2021).

5 Results

This section describes the main findings from our analysis. We first analyze the effect of treatment on beliefs about Russian guilt and support for policy responses. We show results from aggregate analyses, examining the average effect of treatment on beliefs about Russian guilt and support for various policy responses. We then show how heterogeneous treatment effects are consistent with our theoretical predictions (Hypothesis 1). We then analyze the effect of treatment on perceptions of the information source (Hypothesis 2). We again show aggregate results and heterogeneous effects.

5.1 Treatment effects on aggregate posterior beliefs and policy support

Did accusations by the ICC and the United States influence aggregate opinions about Russian guilt and possible governmental responses? The left pane of Figure 4 shows the effect of the U.S. and ICC treatment on agreement with the statement that Russia committed war crimes. For these estimates, we regressed (OLS) responses to the question about whether Russia committed war crimes on an indicator for which treatment the respondent received. The estimates compare a particular treatment group to the control group, excluding the other treatment group.[44] Aggregate treatment effects are generally modest. The aggregate effect of the U.S. treatment is actually *negative*. Overall, the U.S. accusation moved respondents' beliefs in the unintended direction. The ICC treatment effect is also negative, but it is very close to zero.

The right pane of Figure 4 describes the difference in the two treatment effects.[45] While the ICC treatment did not generate a statistically significant change on its own, public belief in Russian war crimes under the ICC treatment is at least significantly less prone to backlash compared to the U.S. treatment. The right panel of the figure illustrates this contrast, showing that beliefs about Russian war crimes are stronger among those exposed to the ICC treatment than among those who received the U.S. treatment.

Figure 5 shows the effects of treatment on the downstream policy responses: support for non-

[44] These specifications have a single intercept. Results are very similar with country-specific intercepts. See appendix.

[45] These estimates describe the effect of the ICC treatment, excluding control observations, i.e. they compare the ICC and U.S. treatment groups.

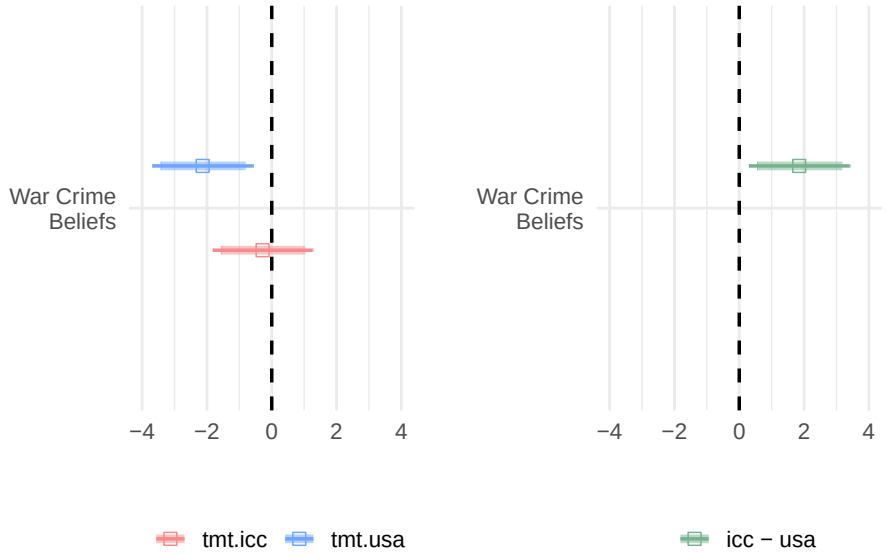


Figure 4: Aggregate treatment effects on posterior beliefs about Russian war crimes.

military aid to Ukraine, military aid, and the sanctions regime.[46] The results are very similar, with the US treatment having negative effects and the ICC treatment having positive effects. The differences in the two treatment effects are also similar.

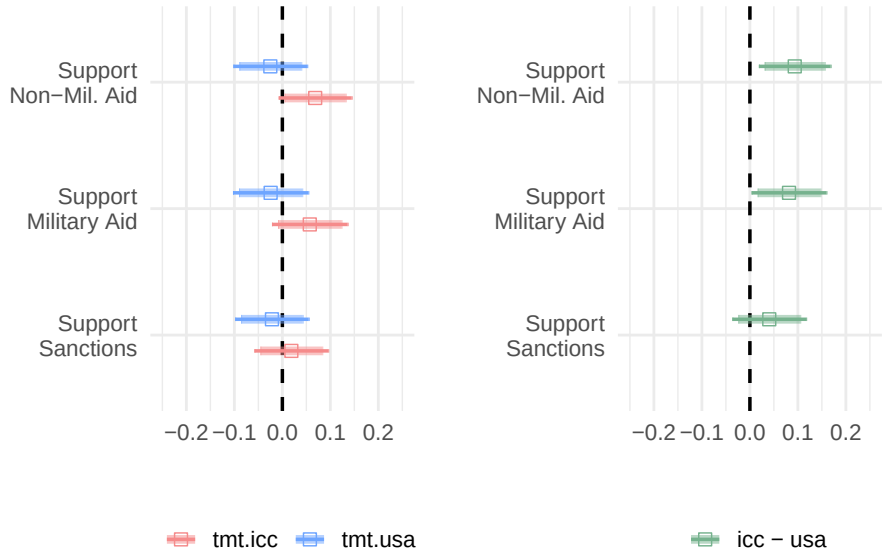


Figure 5: Aggregate treatment effects on support for policy responses.

[46]The estimates are from the same regressions as above, just with different outcome measures.

5.2 Hypothesis 1 Results

Why did the U.S. treatment have negative effects and the ICC treatment have more positive effects? Which respondents were most affected by treatment? The distributions of trust in each source across respondents gives the first clue. Figure 6 shows the smoothed distributions of responses to the question of trust in the United States and ICC, by country. The vertical lines show the sample means. In all four countries, the ICC is viewed as much more trustworthy than the United States. The largest difference was in Turkey, where trust in the ICC was 54.6 compared to 36.3 for the United States, a gap of 18.3 points ($p < 0.001$). In Indonesia, trust in the ICC was 59.3 and trust in the United States was 43.9, a difference of 15.4 points ($p < 0.001$). In South Africa, trust in the ICC was 62.6 compared to 52.5 for the United States, a gap of 10.2 points ($p < 0.001$). The smallest difference was in India, where trust in the ICC was 72.8 and trust in the United States was 67.7, a difference of 5.1 points ($p < 0.001$).

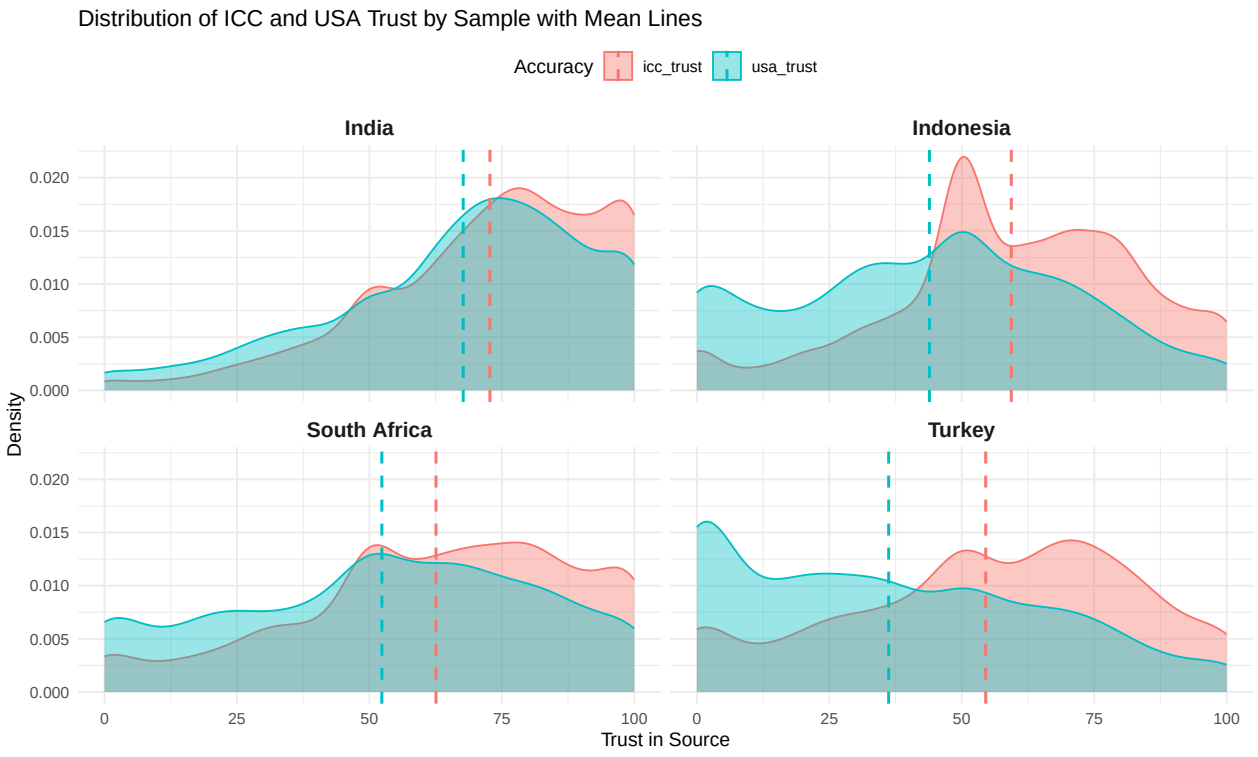


Figure 6: Distribution of trust in sources by country.

To assess Hypotheses 1a and 1b we classified respondents according to whether they were above

or below the sample medians for the measures of prior likelihood of Russian guilt and pre-treatment measures of the accuracy of a source of information.[47] We then estimated the effect of the US and ICC treatments for the subsamples based on above/below median for priors and above/below median based on measures of source accuracy.[48]

Hypothesis 1a predicts the greatest persuasive effects for respondents with low prior probabilities of Russian guilt and high accuracy measures of the source. Hypothesis 1b predicts backlash among respondents with high prior probabilities of Russian guilt and low beliefs about accuracy of the source.

Figure 7 shows the ICC treatment effects by subgroup. The vertical axis shows whether the respondent was above or below the median in their prior beliefs about Russian guilt. The horizontal axis shows whether the respondent was above or below the median in their prior beliefs about the accuracy of the ICC. In other words, its layout matches that of Figure 1. Each cell shows the estimated treatment effect and the p value for a test of whether the treatment effect is different from zero.

Respondents in the bottom right cell – with priors that Russia was innocent and who also had higher prior trust in the ICC were more persuaded by the treatment. Their posterior beliefs about Russian guilt were approximately 3% higher than their priors about Russian guilt. In this cell, for respondents in the control group, the outcome measure for Russian guilt was approximately 62 out of 100. In the treatment group, this outcome measure was approximately 65. Respondents in the top left cell – who thought Russia was guilty but did not trust the ICC – moved their beliefs in the opposite direction, as expected. They lowered their posterior probability of Russian guilt by approximately 3%, from 77 to 74. This pattern matched that predicted by the theoretical model and Hypothesis 1. It is most important to establish that the treatment effects in the top left and bottom right are different from one another. The diagonal arrow shows the p-value for a statistical test of whether the top left effect (backlash group) differs from the bottom right effect (persuasion group). It indicates that the treatment effects differ significantly between the two groups at the p = 0.01 level.

These results are especially striking because, recall, that the aggregate effects were near zero and

[47]The appendix shows the sample sizes for each box and alternate specifications. We used the country-specific medians, but results are nearly identical with global medians. See appendix.

[48]We interacted treatment with indicator variables for which cell a respondent was in, including cell-specific intercepts. Standard errors and test statistics were calculated with the *emmeans* package in R.

insignificant. Those aggregate analyses obscured substantial heterogeneity in treatment effects - heterogeneity that closely matched that predicted by our model. For a respondent in the top left, the ICC treatment lowered her agreement with the statement that Russia was guilty by over three points. For a respondent in the bottom right, treatment persuaded her and increased her belief that Russia was guilty by almost three points.

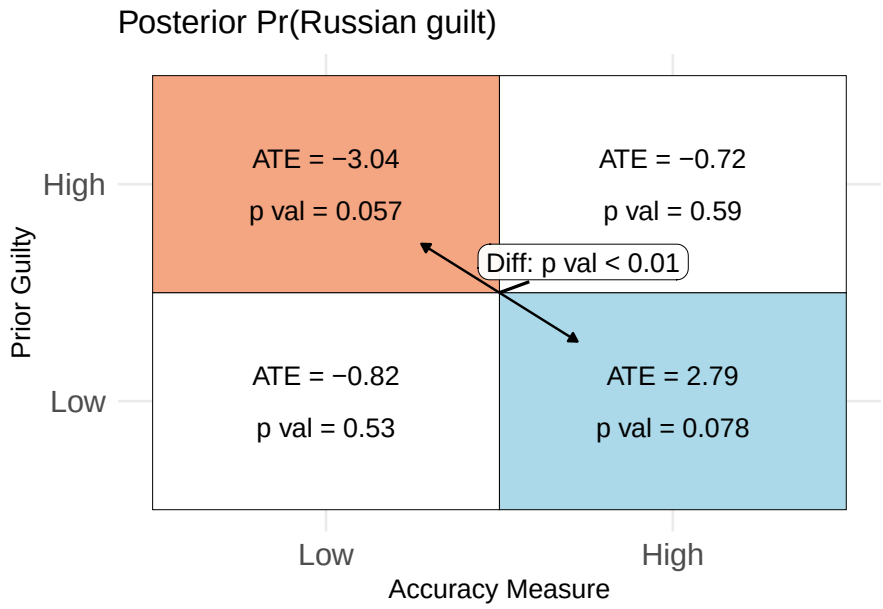


Figure 7: **Effect of ICC treatment on posteriors of Russian guilt.** The ATE captures the effect of treatment among respondents within each subgroup by prior beliefs about Russia and the ICC. The diagonal arrow shows the difference in treatment effects between the backlash and persuasion groups predicted by our model.

Figure 8 shows the ICC treatment effects, in this same way, for all four outcome variables. The top left pane, matches Figure 7. The other panes show treatment effects for sanctions on Russia, non-military aid for Ukraine, and military aid for Ukraine. The patterns are similar. Respondents in the bottom right are those most moved to support sanctions or aid to Ukraine. Backlash is less common, though it tends to be among respondents in the top left. For all four outcome measures, the diagonal arrow shows the treatment effects differ significantly between the backlash and persuasion groups at conventional significance levels. These differences show that respondents in different subgroups react differently to the same information, in a way that generally aligns with our theoretical model

Figure 9 shows the same thing for the U.S. treatment. There are important similarities and dif-

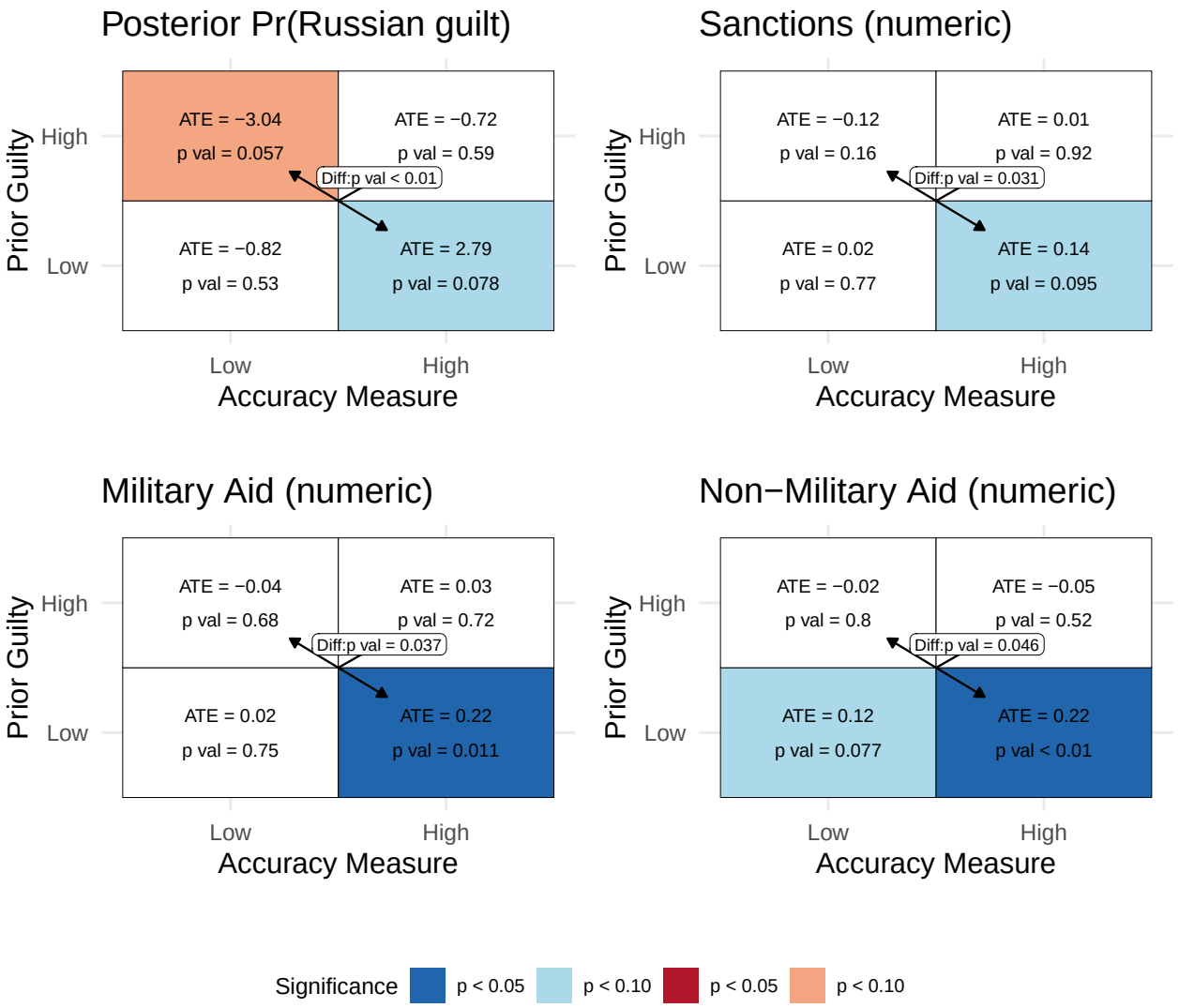


Figure 8: **Effect of ICC treatment on posteriors of all four outcome measure.** The ATE captures the effect of treatment among respondents within each subgroup by prior beliefs about Russia and the ICC. The diagonal arrow shows the difference in treatment effects between the backlash and persuasion groups predicted by our model.

ferences. Backlash is much more prevalent for the U.S. treatment. For respondents in the top left cell, U.S. accusations lowered their posterior beliefs that Russia was guilty by over 4%. In most cells, for all four outcome measures, the U.S. treatment has a negative effect lowering posteriors of Russian guilt or lowering support for policy responses against Russia. Such backlash effect also tends to appear in top-left panel with low perception of U.S accuracy and strong prior beliefs about Russia as predicted by our model. In terms of differences between persuasion and backlash groups, the treatment effects in the top left are statistically different from those in the bottom right for two of the four outcomes—Russian guilt and support for non-military aid. These differences show that respondents in different subgroups move in opposite directions in respond to the identical treatment. For military aid and sanctions, however, the differences are not statistically significant.[49]

These results are also evidence that the effects are as predicted by the theoretical model and not simply floor and ceiling effects. Floor and ceiling effects would show red effects on the top row and blue effects on the bottom row. But this is not the observed pattern. Beliefs and support for policies are not simply being moved away from floors or ceilings. They are being moved in ways that depend *jointly* on prior beliefs and perceptions of the source.

[49] One possible explanation for this null finding is the varying costs of foreign policy options. Military aid and sanctions may be perceived as too costly in this context, even in the face of an accusation. While exploring perceptions of these costs is beyond the scope of this paper, it remains an important question for future research.

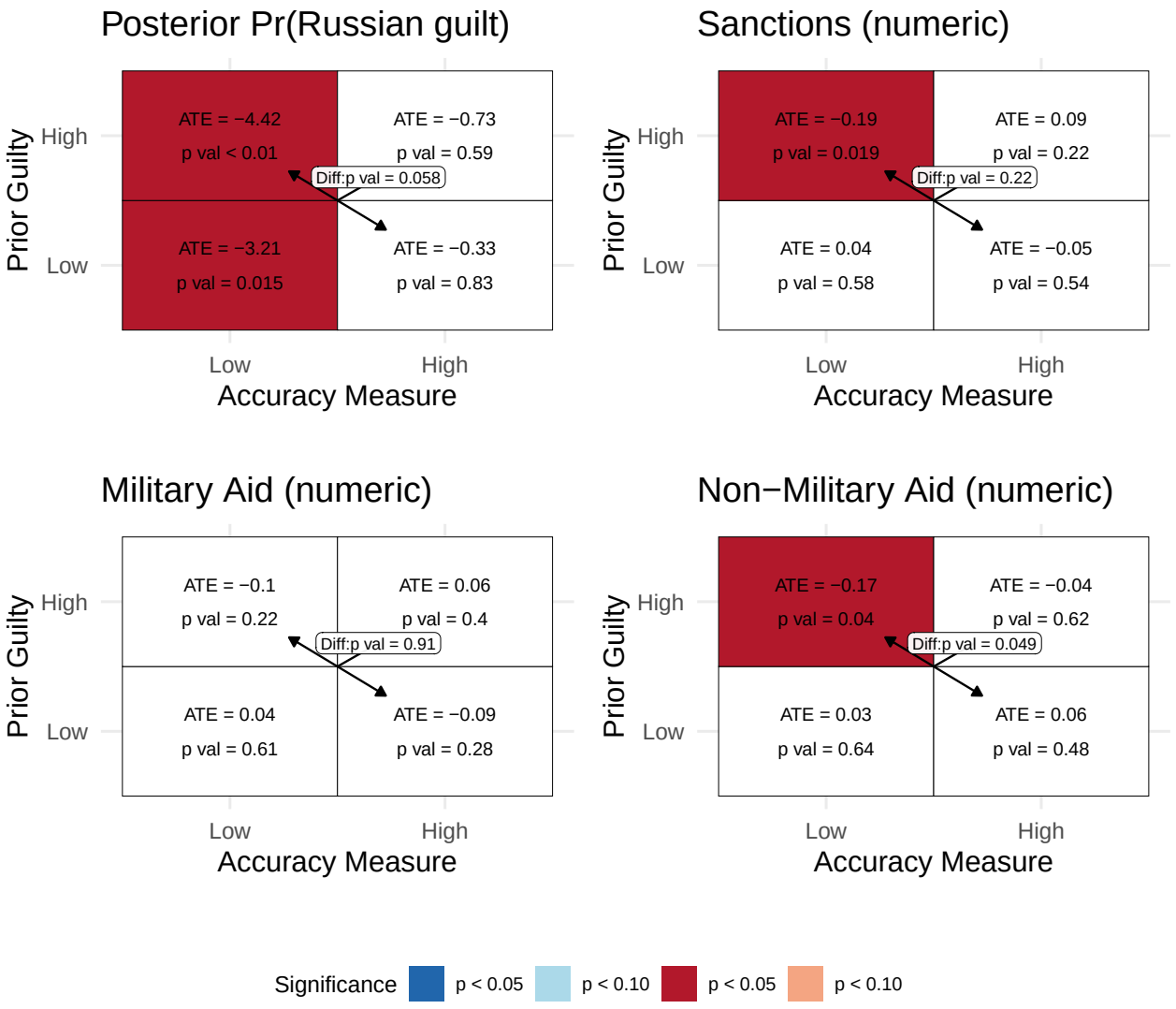


Figure 9: **Effect of USA treatment on posteriors of all four outcome measures** The ATE captures the effect of treatment among respondents within each subgroup by prior beliefs about Russia and the US. The diagonal arrow shows the difference in treatment effects between the backlash and persuasion groups predicted by our model.

5.3 Aggregate Effects on Perceptions of Information Sources

We again look first at the aggregate effect of treatment on perceptions of the trustworthiness of each source. Figure 10 shows results for these outcome variables. When accusations of war crimes were issued by the United States, respondents' trust in the U.S. declined. It reinforced perceptions of the United States as a biased source of information. Trust in the U.S. decreased by approximately 1.5 points on a 100-point scale, a decline of about 4.2%.

In contrast, when the exact same accusation came from the ICC, public trust in the ICC increased, strengthening beliefs in the ICC's impartiality. Trust in the ICC rose by approximately 1.8 points, corresponding to a 4.3% increase. The ICC's signal strengthened public trust of the source.[50]

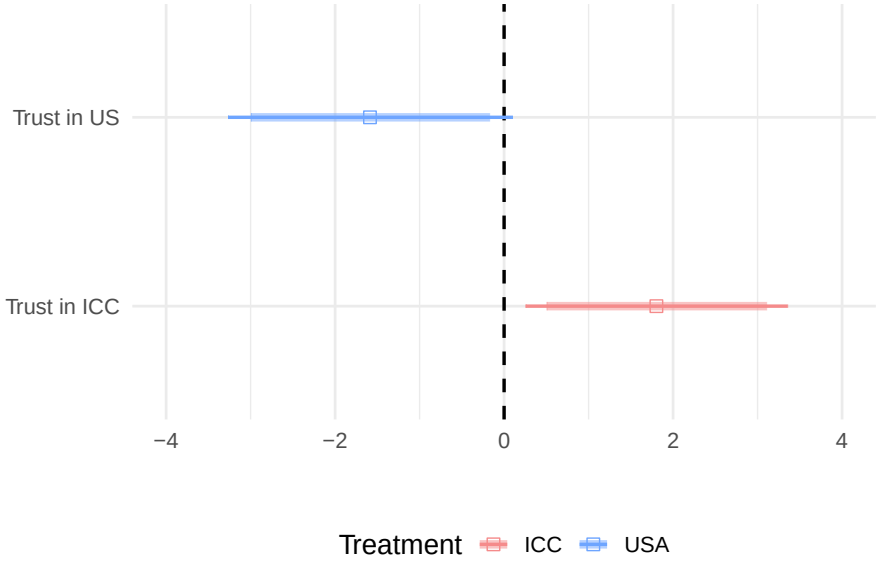


Figure 10: Effect of treatment on trustworthiness of source

5.4 Hypothesis 2 Results

For Hypothesis 2, the results are consistent with the prediction that treatment effects are increasing in priors about Russian guilt. Figure 11 shows how treatment effects vary with prior beliefs about Russian guilt. These are estimates from a linear interaction term model, interacting treatment with priors

[50] We also tested whether treatment affected perceptions of ICC legitimacy, as a related outcome measure. Treatment increased perceptions of ICC legitimacy. See appendix.

about Russian guilt. As expected, the lines are upward sloping. When the source gives information that matches the respondent's priors, the respondent increases their trust in the source.

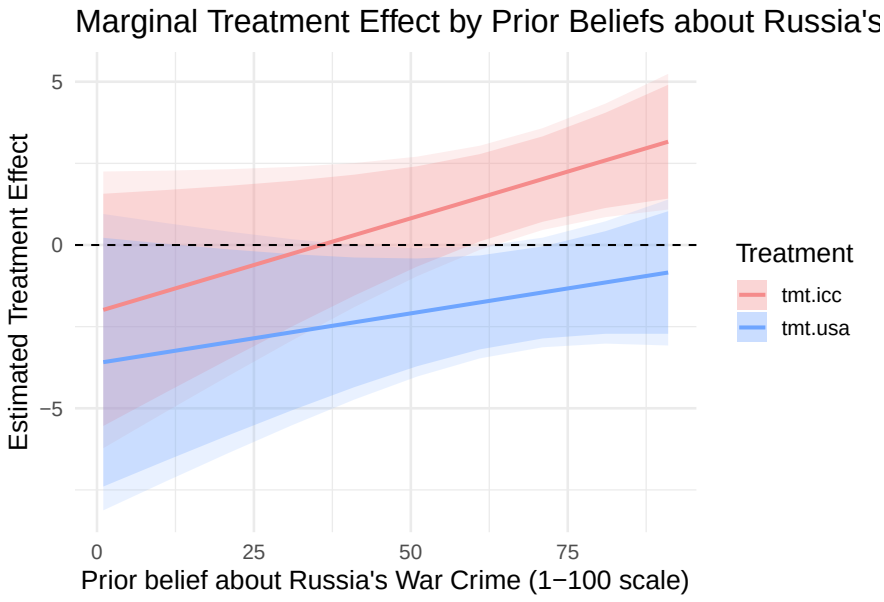


Figure 11: Effect of treatment on posteriors about source quality, as prior beliefs vary.

However, the differences between the results and other parts of the model's theoretical predictions are also interesting. The positive aggregate effect of the ICC treatment, compared to the negative aggregate effect for the United States is surprising (as seen in Figure 10). Similarly, the estimated treatment effects of the ICC are more positive/less negative for respondents all along the range of prior beliefs about Russian guilt (as see in Figure 11). Even for respondents that believe strongly in Russian guilt *ex ante* still have negative estimated treatment effects for the U.S. treatment. Respondents generally started with higher prior beliefs about the accuracy of the ICC. The model would have suggested that treatment effects would be *weaker* for the ICC compared to the United States.

Figure 12 shows 2x2 style boxes that we used to estimate treatment effects on posterior beliefs about Russian guilt. We again have priors about accuracy on the horizontal axis and priors about Russian guilt on the vertical axis. The outcome variable is now posteriors beliefs about a source's trustworthiness. We would expect the largest, positive effects to be in the top left of each figure. This is true for both the ICC and the United States. For the ICC, the respondents who had low initial trust in the Court but then were treated with information that matched their priors, showed a significant

increase in trust in the ICC. The only positive treatment effects for the United States were also found in the upper left quadrant, though these effects were insignificant. We would also expect the largest negative effects in the bottom right, which is true for the U.S. treatment. Respondents in this quadrant had lower post-treatment views of U.S. credibility.

The model predicts that treatment effects should be more positive for respondents who started with lower prior assessments of the source's accuracy. The treatment effects for the ICC are positive in all four quadrants and the U.S. treatment effects are negative in three out of four quadrants. For most respondents, even when the ICC gives a signal that contradicts their priors about Russian guilt, they are still increasing their perceptions of the ICC's trustworthiness. Similarly, even when the United States sends a signal that matches the respondent's priors, they generally downgrade their beliefs about the United States as a trustworthy source of information.

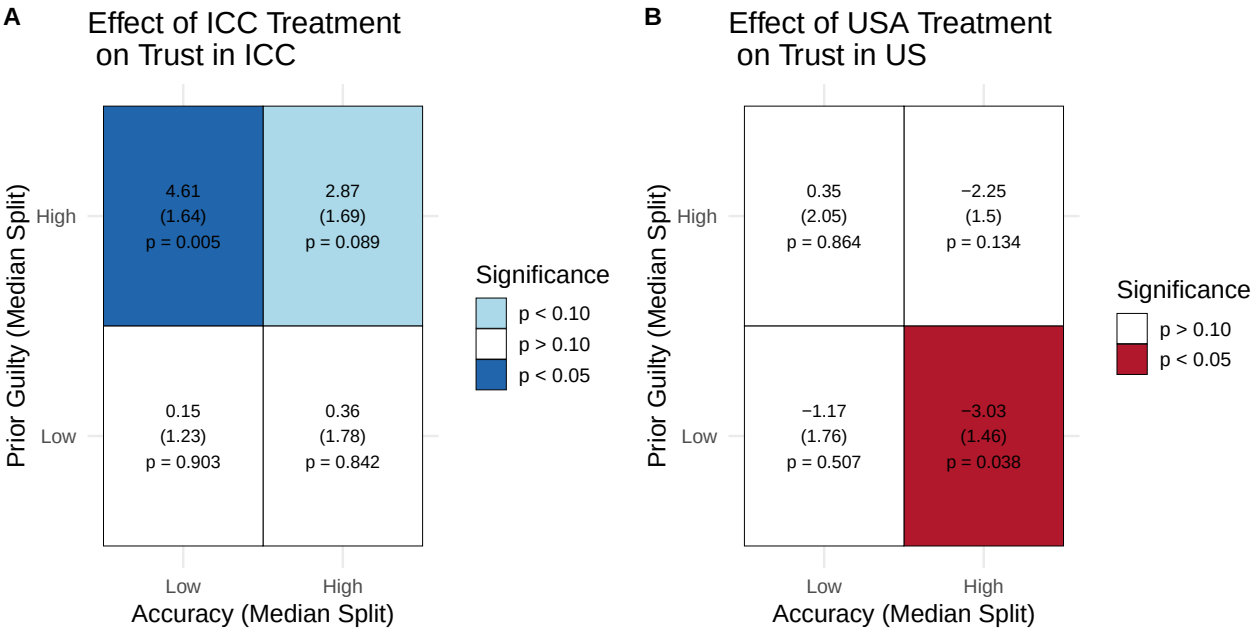


Figure 12: Effect of treatment on beliefs about the source, by prior beliefs about Russia and the source

The aggregate results – with treatment improving views of the ICC's trustworthiness and decreasing views of the United States' trustworthiness – is striking, because it shows respondents diverging in their beliefs about source trustworthiness, despite both sources sending the same signal. Most models of persuasion generally predict convergence, when two sources say the same thing.[51]

[51]For a more extended discussion of the conditions under which convergence in beliefs about the sender may not

This difference between the theoretical model's prediction and the findings is especially interesting because it suggests that there is something distinct about the signalling advantage of the ICC over the United States. Despite giving the same signal, and despite the deck being stacked against finding positive effects of the ICC treatment, the Court is *still* better able to use signals to persuade audiences about its credibility. The skepticism that respondents held towards the United States, *ex ante*, was reinforced and deepened, even though our treatment had the United States give respondents the same piece of information as the information given by ICC.

One possible explanation is that respondents infer different information from the treatment, even though the wording is identical, and the post-treatment measure of trust elicits something about this additional information. When the ICC accuses Russia of war crimes, it is possible that this conveys information that a legal body has evaluated evidence following a particular legal procedure and come to a corresponding conclusion about war crimes. Perhaps this conveys the idea of an investigation, with evidence weighed and debated in open court. And the Court only sends its signal that Russia has committed war crimes after this careful process. When the United States accuses Russia of war crimes, it is possible that the respondent does not infer anything about deliberation or weighing of evidence.

If the post-treatment measurement of trust captures more than just a posterior belief about the accuracy of a signal – i.e. it also captures beliefs about the quality of the process to arrive at that signal, beyond the statement of the signal itself – then that could explain this unexpected aspect of the treatment effects. The treatment effect is therefore capturing an updated view about the source's process, not just that the source's output was a signal that conveys the right answer about the state of the world. If so, that could explain this unexpected aspect of treatment effects. To assess these possibilities, we would need different outcome measures that captured these potential unintended effects.

converge, see Acemoglu, Chernozhukov, and Yildiz (2016) and Cheng and Hsiaw (2022).

6 Conclusions

Accusations about violations of international law generate both persuasion and backlash among publics in global swing states. Whether accusations persuade or backfire depends on who sends the message and whom they are trying to convince. Using survey experiments in four swing states – India, Indonesia, Turkey, and South Africa – we show that identical accusations against Russia lead to divergent reactions when attributed to different sources. Accusations from the International Criminal Court produce modest persuasion, particularly among those who trust the ICC and do not already hold strong prior beliefs about Russian guilt. In contrast, accusations from the United States often backfire, undermining both belief in the accusation and trust in the sender. We offer a theoretical model to explain these patterns, showing how belief updating is jointly shaped by priors and perceived source credibility.

Our results offer cautious optimism about the potential for international organizations to build persuasive power through consistent and credible engagement. All but one of our surveyed countries has refused to join the Court. Yet, even a Court that does not have universal support, especially among global swing states, was able to persuade some subsets of respondents. Even more encouragingly, its messages also increased perceptions of the Court's own trustworthiness and legitimacy, even among respondents that doubted the Court's message. A hopeful aspect of this finding is that the Court may be building a well of legitimacy that makes it even more persuasive in the future. The Court is potentially building its well of legitimacy, despite many of its decisions being met with disagreement or ambivalence.

However, the contrast between the ICC findings and those for the United States are ominous with respect to U.S. credibility. Our results suggest that the United States' messaging is more harmful for its agenda than remaining silent, at least with respect to public opinion in critical swing states. U.S. messaging was less effective even than the often-maligned ICC, and it backfired for a plurality of respondents. Notably, our surveys were implemented before the 2024 U.S. Presidential elections and global perceptions of the United States have further declined after the survey. According to surveys conducted in over 100 countries, China's net favorability rating is now 19 points higher than the

United States' and Russia's net rating is only 4 points behind that of the U.S.![52] The ICC can at least hope that its messaging triggers a positive feedback loop, where accusations enhance credibility which makes future accusations more effective. The United States, on the other hand, needs to worry about a doom spiral, where its lack of credibility causes accusations to backfire and make its future messaging less credible. U.S. policymakers that discount the importance of soft power would do well to remember that credibility helps persuade others to back concrete punishments for U.S. adversaries, like sanctions or arms transfers to allies.

Our arguments speak to a broader literature on how foreign sources of information shape public opinion. These signals may come from international organizations, public diplomacy efforts, or other states' naming and shaming. We demonstrate a tractable framework that generates testable predictions across contexts. This framework makes it clear how aggregate effects can obscure important heterogeneity. Our results show that such messages can either persuade or provoke backlash, depending on who delivers the message and how audiences evaluate the credibility of the sender.

To evaluate the generalizability of these expectations, it is essential to examine whether similar patterns of persuasion and backlash arise in other settings. For instance, in 2024, the ICC issued arrest warrants for both Hamas and Israeli leaders in connection with the Israel–Palestine conflict. The Court's involvement in this case sparked significant public debate and controversy across countries and political groups. Some governments, especially member states of the Rome Statue, welcomed the investigation, while other governments expressed strong opposition.[53] The United States, for example, condemned the Court's actions and ultimately imposed sanctions on the ICC in 2025.[54] Examining audience responses in different cases, where priors, political alignments, and trust in information sources vary, would provide an important test of our theoretical expectations.

Future research can also extend the analysis by varying the type of messenger, including both international organizations and individual states. Institutions such as the International Court of Justice, the United Nations, and the European Union also seek to shape public opinion through the dissem-

[52] Nira Democracy Perception Index 2025 Report. https://www.niradata.com/dpi.

[53] https://www.justsecurity.org/105064/arrest-warrants-state-reactions-icc/.

[54] https://www.whitehouse.gov/presidential-actions/2025/02/imposing-sanctions-on-the-international-criminal-court/.

ination of information. Whether such signals from these actors elicit similar patterns of persuasion or backlash represents a valuable avenue for future studies. Likewise, individual states engage in efforts to influence foreign publics— campaigns described as public diplomacy or propaganda. Testing our theoretical expectations across a range of senders of information would offer an evaluation of the theory's external validity.

7 References

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz. 2016. "Fragility of Asymptotic Agreement Under Bayesian Learning." *Theoretical Economics* 11 (1): 187–225.

Anjum, Gulnaz, Adam Chilton, and Zahid Usman. 2021. "United Nations Endorsement and Support for Human Rights: An Experiment on Women's Rights in Pakistan." *Journal of Peace Research* 58 (3): 462–78.

Bassan-Nygate, Lotem, Jonathan Renshon, Jessica LP Weeks, and Chagai M Weiss. 2024. "The Generalizability of IR Experiments Beyond the United States." *American Political Science Review*, 1–16.

Bearce, David H, and Thomas R Cook. 2018. "The First Image Reversed: IGO Signals and Mass Political Attitudes." *The Review of International Organizations* 13 (4): 595–619.

Brutger, Ryan. 2021. "The Power of Compromise: Proposal Power, Partisanship, and Public Support in International Bargaining." *World Politics* 73 (1): 128–66.

Brutger, Ryan, and Anton Strezhnev. 2022. "International Investment Disputes, Media Coverage, and Backlash Against International Law." *Journal of Conflict Resolution* 66 (6): 983–1009.

Burcu Bayram, A, Eric Keels, and Efe Tokdemir. 2024. "Enforcement of International Human Rights Law: A Comparative Exploration of Alternative Public Opinion Channels." *The British Journal of Politics and International Relations*, 13691481241305975.

Búzás, Zoltán I, and Lotem Bassan-Nygate. 2024. "Race, Shaming, and International Human Rights." *American Journal of Political Science*.

Carnegie, Allison, Richard Clark, and Lisa Fan. 2024. "Multilateral Messaging: International Organizations, Populism, and Social Media."

Chapman, Terrence L. 2007. "International Security Institutions, Domestic Politics, and Institutional Legitimacy." *Journal of Conflict Resolution* 51 (1): 134–66.

Chapman, Terrence L, and Stephen Chaudoin. 2020. "Public Reactions to International Legal Institutions: The International Criminal Court in a Developing Democracy." *The Journal of Politics* 82 (4): 1305–20.

Chaudoin, Stephen. 2014. "Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions." *International Organization* 68 (1): 235–56.

———. 2023. "How International Organizations Change National Media Coverage of Human Rights." *International Organization* 77 (1): 238–61.

Chaudoin, Stephen, Brian J Gaines, and Avital Livny. 2021. "Survey Design, Order Effects, and Causal Mediation Analysis." *The Journal of Politics* 83 (4): 1851–56.

Cheng, Haw, and Alice Hsiaw. 2022. "Distrust in Experts and the Origins of Disagreement." *Journal of Economic Theory* 200: 105401.

Chivvis, Christopher S., Elina Noor, and Beatrix Geaghan-Breiner. 2023. "Indonesia in the Emerging World Order." 2023. https://carnegieendowment.org/research/2023/11/indonesia-in-the-emerging-world-order?lang=en.

Choi, Ha Eun, JiHwan Jeong, Amanda Murdie, Byungwon Woo, and Hyunjin Yim. 2023. "UN Secretary-General Visits and Human Rights Diplomacy." In *Paper Presented at the 15th Annual Conference Political Economy of International Organization*.

Chow, Wilfred M, and Dov H Levin. 2024. "The Diplomacy of Whataboutism and US Foreign Policy Attitudes." *International Organization* 78 (1): 103–33.

Chu, Jonathan Art. 2025. *Social Cues: How the Liberal Community Legitimizes Humanitarian War*. Cam-

bridge University Press.
Chung, Seowoo. 2025. "Strategic Censorship? Public Opinion, Authoritarian Politics, and the International Trade Regime."
Cohen, Harlan, and Ryan Powers. 2024. "Judicialization and Public Support for Compliance with International Commitments." *International Studies Quarterly* 68 (3): sqae078.
Cope, Kevin L. 2023. "Measuring Law's Normative Force." *Journal of Empirical Legal Studies* 20 (4): 1005–44.
Cope, Kevin L, and Charles Crabtree. 2020. "A Nationalist Backlash to International Refugee Law: Evidence from a Survey Experiment in Turkey." *Journal of Empirical Legal Studies* 17 (4): 752–88.
Coppock, Alexander. 2023. *Persuasion in Parallel: How Information Changes Minds about Politics*. University of Chicago Press.
Davis, Christina, and Taegyun Lim. 2025. "Targeted Sanctions and Maritime Shipping: Evidence from Satellite Tracking of Vessels."
Delavande, Adeline. 2014. "Probabilistic Expectations in Developing Countries." *Annu. Rev. Econ.* 6 (1): 1–20.
Dellmuth, Lisa M, and Jonas Tallberg. 2021. "Elite Communication and the Popular Legitimacy of International Organizations." *British Journal of Political Science* 51 (3): 1292–1313.
Ecker-Ehrhardt, Matthias, Lisa Dellmuth, and Jonas Tallberg. 2024. "Ideology and Legitimacy in Global Governance." *International Organization* 78 (4): 731–65.
Efrat, Asif, and Omer Yair. 2023. "International Rankings and Public Opinion: Compliance, Dismissal, or Backlash?" *The Review of International Organizations* 18 (4): 607–29.
Fontaine, Richard, and Daniel M Kliman. 2013. "International Order and Global Swing States." *The Washington Quarterly* 36 (1): 93–109.
Fontaine, Richard, and Gibbs McKinley. 2025. "Global Swing States and the New Great Power Competition." *The Washington Quarterly* 48 (2): 7–28.
Gentzkow, Matthew, Michael B Wong, and Allen T Zhang. Forthcoming. "Ideological Bias and Trust in Information Sources." *American Economic Journal: Microeconomics*, Forthcoming.
Ghassim, Farsan. 2024. "Effects of Self-Legitimation and Delegitimation on Public Attitudes Toward International Organizations: A Worldwide Survey Experiment." *International Studies Quarterly* 68 (2): sqae012.
Goldsmith, Benjamin E, and Yusaku Horiuchi. 2009. "Spinning the Globe? US Public Diplomacy and Foreign Public Opinion." *The Journal of Politics* 71 (3): 863–75.
Goldsmith, Benjamin E, Yusaku Horiuchi, and Kelly Matush. 2021. "Does Public Diplomacy Sway Foreign Public Opinion? Identifying the Effect of High-Level Visits." *American Political Science Review* 115 (4): 1342–57.
Grieco, Joseph M, Christopher Gelpi, Jason Reifler, and Peter D Feaver. 2011. "Let's Get a Second Opinion: International Institutions and American Public Support for War." *International Studies Quarterly* 55 (2): 563–83.
Hansen, Ben B, and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science*, 219–36.
Helfer, Laurence R, and Anne E Showalter. 2017. "Opposing International Justice: Kenya's Integrated Backlash Strategy Against the ICC." *International Criminal Law Review* 17 (1): 1–46.
Keck, Margaret E, and Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Cornell University Press.
Kertzer, Joshua D, Brian C Rathbun, and Nina Srinivasan Rathbun. 2020. "The Price of Peace: Mo-

tivated Reasoning and Costly Signaling in International Relations." *International Organization* 74 (1): 95–118.

Kliman, Daniel M. 2012. "The West and Global Swing States." *The International Spectator* 47 (3): 53–64.

Lee, James. 2022. "Foreign Aid, Development, and US Strategic Interests in the Cold War." *International Studies Quarterly* 66 (1).

Little, Andrew T. 2025. "How to Distinguish Motivated Reasoning from Bayesian Updating." *Political Behavior*, 1–25.

Lupia, Arthur, and Mathew D McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge University Press.

Mattingly, Daniel, Trevor Incerti, Changwook Ju, Colin Moreshead, Seiki Tanaka, and Hikaru Yamagishi. 2024. "Chinese State Media Persuades a Global Audience That the 'China Model' Is Superior: Evidence from a 19-Country Experiment." *American Journal of Political Science*.

Mattingly, Daniel, and James Sundquist. 2023. "When Does Public Diplomacy Work? Evidence from China's 'Wolf Warrior' Diplomats." *Political Science Research and Methods* 11 (4): 921–29.

Mikulaschek, Christoph. 2023. "The Responsive Public: How European Union Decisions Shape Public Opinion on Salient Policies." *European Union Politics* 24 (4): 645–65.

Mikulaschek, Christoph, and Michal Parizek. 2025. "How Media Coverage Shapes the Effect of IOs on Public Attitudes: Quasi-Experimental Evidence on Mass Opinion about Russia's Leadership in 49 Countries."

Morrison, Kelly. 2024. "Named and Shamed: International Advocacy and Public Support for Repressive Leaders." *Journal of Conflict Resolution* 68 (2-3): 294–321.

Morse, Julia C, and Tyler Pratt. 2022. "Strategies of Contestation: International Law, Domestic Audiences, and Image Management." *The Journal of Politics* 84 (4): 2080–93.

———. 2025. "Smoke and Mirrors: Strategic Messaging and the Politics of Noncompliance."

Murdie, Amanda M, and David R Davis. 2012. "Shaming and Blaming: Using Events Data to Assess the Impact of Human Rights INGOs." *International Studies Quarterly* 56 (1): 1–16.

Pauselli, Gino. 2023. "Look Who Is Talking: Direct and Indirect Effects of Criticism on LGBT Rights." *Available at SSRN 4317082*.

Recchia, Stefano, and Jonathan Chu. 2021. "Validating Threat: IO Approval and Public Support for Joining Military Counterterrorism Coalitions." *International Studies Quarterly* 65 (4): 919–28.

Rhee, Kasey, Charles Crabtree, and Yusaku Horiuchi. 2024. "Perceived Motives of Public Diplomacy Influence Foreign Public Opinion." *Political Behavior* 46 (1): 683–703.

Spilker, Gabriele, Quynh Nguyen, and Thomas Bernauer. 2020. "Trading Arguments: Opinion Updating in the Context of International Trade Agreements." *International Studies Quarterly* 64 (4): 929–38.

Suong, Clara H, Scott Desposato, and Erik Gartzke. 2024. "Ubiquitous but Heterogeneous: International Organizations' Influence on Public Opinion in China, Brazil, Japan, and Sweden." *International Relations of the Asia-Pacific*, lcae018.

Syropoulos, Constantinos, Gabriel Felbermayr, Aleksandra Kirilakha, Erdal Yalcin, and Yoto V Yotov. 2024. "The Global Sanctions Data Base–Release 3: COVID-19, Russia, and Multilateral Sanctions." *Review of International Economics* 32 (1): 12–48.

Terman, Rochelle. 2023. "The Geopolitics of Shaming: When Human Rights Pressure WorksâÂ Â and When It Backfires."

Thompson, Alexander. 2006. "Coercion Through IOs: The Security Council and the Logic of Infor-

mation Transmission." *International Organization* 60 (1): 1–34.

———. 2015. *Channels of Power: The UN Security Council and US Statecraft in Iraq.* Cornell University Press.

Voeten, Erik. 2020. "Populism and Backlashes Against International Courts." *Perspectives on Politics* 18 (2): 407–22.

Wang, Austin Horng-En, Charles KS Wu, Yao-Yuan Yeh, and Fang-Yu Chen. 2023. "High-Level Visit and National Security Policy: Evidence from a Quasi-Experiment in Taiwan." *International Interactions* 49 (1): 132–46.

Zvobgo, Kelebogile. 2019. "Human Rights Versus National Interests: Shifting US Public Attitudes on the International Criminal Court." *International Studies Quarterly* 63 (4): 1065–78.

A Theory

A.1 Expression for treatment effect about state of the world

The expression for the treatment effect for posteriors about the state of the world is a straightforward application of Bayes' rule. We omit the i subscripts for simplification.

$$\Pr(S = 1 \mid s_1) = \frac{\Pr(S = 1)\Pr(s_1 \mid S = 1)}{\Pr(S = 1)\Pr(s_1 \mid S = 1) + (1 - \Pr(S = 1))\Pr(s_1 \mid S = 0)}$$

We can compute the likelihood terms using expectations under the Beta distribution:

$$\Pr(s_1 \mid S = 1) = \int_0^1 f(\sigma) \cdot \sigma \, d\sigma = \mathbb{E}[\sigma] = \frac{\alpha}{\alpha + \beta}$$

$$\Pr(s_1 \mid S = 0) = \int_0^1 f(\sigma) \cdot (1 - \sigma) \, d\sigma = 1 - \mathbb{E}[\sigma] = \frac{\beta}{\alpha + \beta}$$

Substituting into Bayes' rule:

$$\Pr(S = 1 \mid s_1) = \frac{\Pr(S = 1) \cdot \frac{\alpha}{\alpha+\beta}}{\Pr(S = 1) \cdot \frac{\alpha}{\alpha+\beta} + (1 - \Pr(S = 1)) \cdot \frac{\beta}{\alpha+\beta}}$$

Simplifying:

$$\Pr(S = 1 \mid s_1) = \frac{\Pr(S = 1) \cdot \alpha}{\Pr(S = 1) \cdot \alpha + (1 - \Pr(S = 1)) \cdot \beta}$$

Letting $\pi = \Pr(S = 1)$:

$$\Pr(S = 1 \mid s_1) = \frac{\pi\alpha}{\pi\alpha + (1 - \pi)\beta}$$

So the treatment effect, with i subscripts reintroduced, is:

$$\Pi_i = \frac{\pi_i\alpha_i}{\pi_i\alpha_i + (1 - \pi_i)\beta_i} - \pi_i$$

A.2 Expression for treatment effect about source accuracy

Recall that the treatment effect for source accuracy is: $\Sigma_i = \mathbb{E}\left[\sigma_i | s_1\right] - \mathbb{E}\left[\sigma_i\right]$.

The first term, omitting i subscripts again, $\mathbb{E}\left[\sigma | s_1\right]$ can be written by breaking down the two possibilities - either the sender was right or they were wrong.

$$\mathbb{E}\left[\sigma | s_1\right] = \Pr(S = 1 \mid s_1) \cdot \mathbb{E}\left[\sigma | s_1, S = 1\right] + \Pr(S = 0 \mid s_1) \cdot \mathbb{E}\left[\sigma | s_1, S = 0\right]$$

Substituting the expression for $\Pr(S = 1 \mid s_1)$ from above...

$$\mathbb{E}\left[\sigma | s_1\right] = \frac{\pi\alpha}{\pi\alpha + (1 - \pi)\beta} \cdot \mathbb{E}\left[\sigma | s_1, S = 1\right] + (1 - \frac{\pi\alpha}{\pi\alpha + (1 - \pi)\beta}) \cdot \mathbb{E}\left[\sigma | s_1, S = 0\right]$$

$$\mathbb{E}\left[\sigma|s_1\right] = \frac{\pi\alpha}{\pi\alpha + (1-\pi)\beta} \cdot \mathbb{E}\left[\sigma|s_1, S=1\right] + \frac{(1-\pi)\beta}{\pi\alpha + (1-\pi)\beta} \cdot \mathbb{E}\left[\sigma|s_1, S=0\right]$$

For the term $\mathbb{E}\left[\sigma|s_1, S=1\right]$, this occurs when the signal sender gets it "right." Their signal correctly matched the state of the world. From the Beta-Binomial conjugacy, their "new" σ is distributed Beta with parameters $\alpha + 1$ and β. The expectation of that new distribution is $\frac{\alpha+1}{\alpha+\beta+1}$. For the term $\mathbb{E}\left[\sigma|s_1, S=0\right]$, this occurs when the signal sender gets it "wrong." The expectation of that new distribution is $\frac{\alpha}{\alpha+\beta+1}$.

Substituting these expressions in…

$$\mathbb{E}\left[\sigma|s_1\right] = \frac{\pi\alpha}{\pi\alpha + (1-\pi)\beta} \cdot \frac{\alpha+1}{\alpha+\beta+1} + \frac{(1-\pi)\beta}{\pi\alpha + (1-\pi)\beta} \cdot \frac{\alpha}{\alpha+\beta+1}$$

Simplifying and re-adding i subscripts…

$$\mathbb{E}\left[\sigma_i|s_1\right] = \frac{\alpha_i}{\alpha_i + \beta_i + 1} \cdot \left[1 + \frac{\pi_i}{\pi_i\alpha_i + (1-\pi_i)\beta_i}\right]$$

Note that this expression is increasing in π_i. By extension, the treatment effect expression is also increasing in π_i.

The full treatment effect expression is…

$$\Sigma_i = \frac{\alpha_i}{\alpha_i + \beta_i + 1} \cdot \left[1 + \frac{\pi_i}{\pi_i\alpha_i + (1-\pi_i)\beta_i}\right] - \frac{\alpha_i}{\alpha_i + \beta_i}$$

A.3 Relation to motivated reasoning models

Plenty of research contrasts Bayesian models of belief updating with alternate models of belief formation, such as those based on motivated reasoning. In motivated reasoning models, individuals form posteriors based on accuracy and directional motives. They may want to get their posteriors "right" (an accuracy motive), but they may also like it when their posteriors are closer to a preferred point (the directional motive). Kertzer, Rathbun, and Rathbun (2020) is a good example from international relations research. They describe how motivated reasoning conditions individuals' reactions to information about costly signalling. "It is precisely those who are motivated to find evidence of a costly signal who act as classic signalling models would expect, while those motivated not to update their beliefs do not respond to the treatments to the same degree, and sometimes not at all" (97). They predict, and find, that individuals with more cooperative internationalist attitudes and/or less militant internationalist attitudes will respond more to costly signals. In their particular application, liberals and those with more positive feelings toward Iran responded more to costly signals from Iran.

Coppock (2023) (ch 7) and Little (2025) both argue that Bayesian models and most motivated reasoning models are indistinguishable with most experimental designs. If a piece of information moves a respondent in a particular way, this could be because she had a particular configuration of priors and accuracy beliefs about the signal *or* because she had particular biases about the direction of her preferred posterior belief. In the example from Kertzer, Rathbun, and Rathbun (2020), those who responded most to a signal may have had directional motives or they may have had different beliefs about the likelihood function generating those signals. A cooperative internationalist may subconsciously think "I am responding to this treatment in the intended direction because it pushes me towards my preferred posterior" *or* they may think "I am responding to this treatment in the intended direction because signals like this are more credible." Source credibility is sometimes described as a likelihood ratio, e.g. (Pr(Iran is peaceful | signal) / Pr(Iran is peaceful | no signal)). Without measurements of priors and the respondent's beliefs about the signal's credibility, these alternatives are impossible to distinguish from one another.

We do not attempt to resolve this voluminous debate about Bayesian models versus their alternatives. Rather, we make two remarks. First, whatever model is used, it should make precise predictions about the direction and magnitude of treatment effects. If those predicted effects are moderated by receiver characteristics (like priors), then the model should make apparent what must be measured pre-treatment to test predictions about who will be most moved by a treatment. Making predictions based on Bayesian *or* motivated reasoning models generally requires measurements of priors and likelihood functions. Coppock argues that we can't tell Bayesian and motivated reasoning stories apart because "We would love to know if changing a likelihood changed a posterior, holding exposure to evidence constant, since that would provide direct evidence for the Bayesian model. But we can't, because likelihood functions are imaginary constructs whose existence in people's minds we can only posit." (137-8). However, just because likelihoods are hard to manipulate, this does not mean that they are impossible to measure. With measurements of priors and likelihoods, Bayes rule gives a predicted posterior, and by extension, a predicted treatment effect, that can be assessed against data. A motivated reasoning model would require those two measurements as well.[55]

[55] Little (2025) shows that these stories will still be indistinguishable, since the priors themselves could be generated from directional motives. We agree. However, our goal again is not to prove or disconfirm the existence of motivated reasoning models. Our goal is to say "conditional on observing priors and likelihoods, Bayes rule gives useful predictions about the types of individuals for whom treatment effects will be largest."

Second, in the context of diplomatic messaging and IO endorsements, it is important for any model of updating to accommodate persuasion *and* backlash. Existing work gives strong reasons to think that both phenomena occur in the real world.[56] Therefore, any model of the effects of diplomatic or IO messaging should be capable of yielding both types of effects. In most applications of motivated reasoning models, backlash does not occur. Predicted treatment effects may be muted, such as when a receiver chooses to discard information that does not match her priors. However, they generally do not generate predictions where information moves receivers in the opposite of its intended direction.

B Complete Survey Instrument

This section of the appendix describes every item on the survey. Researchers have understandably become more worried about mining, particularly when investigating heterogeneous treatment effects - as is the focus of this paper. For readers worried that these heterogeneous treatment effect arguments are an example of mining, we would note that this is the survey in its entirety. The survey was designed solely to assess the predictions of the theoretical model – heterogeneous effects from prior beleifs about the state of the world and the accuracy of information sources – and then contrast them with possible heterogeneous treatment effects based on a commonly used moderator, cooperative internationalism. To that end, the survey is designed to measure the moderating quantities that moderate treatment effects in the theoretical model. There aren't other moderators that we could potentially mine, or at least none that are tied directly to a formal model that makes precise predictions about heterogeneous treatment effects.

We first asked for informed consent. Respondents then had to pass a simple attention check that said "People are very busy these days and many do not have time to follow what goes on in the government. *We are testing whether people read questions.* To show that you've read this much, answer *both*"extremely interested" *and* "very interested."

We then presented the following six blocks, with their order randomized. The first block measured the respondent's prior beliefs that a country had broken international law. The underlined text below was not displayed to respondents. It is only here for readability. We included China and the United States to have other countries, but the key item here was the question about Russia.

- Prior beliefs about breaking international law Countries sometimes violate international laws of war that restrict attacking civilians and other acts. In your opinion, what is the percent chance that the countries below have violated international laws of war over the last 5 years? (0-100, order of items randomized)
  - Russia
  - United States
  - China

The next set of blocks measured pre-treatment views about the accuracy of each source. Since respondents would not have been able to express their answers in terms of a likelihood function (e.g. "What's the probability the ICC says Russia is guilty if they are guilty?"), we used three different types of questions: about whether a country/international organization tells the truth, whether

[56] Eg Goldsmith and Horiuchi (2009) on diplomacy and Terman (2023) on shaming.

they are a trustworthy source of information, and a general feeling thermometer. The key items are those asking about the United States or the ICC. We again included other entities so that the entire focus was not just on the United States and ICC.

- Trustworthiness There are many sources of information about international affairs. Some sources of information are trustworthy and others are not. On a scale of 1-100, with zero being the least trustworthy and 100 being the most trustworthy, where would you place the following sources of information? (order of items randomized)
  - The International Criminal Court
  - The United States government
  - The media

- Countries Telling the Truth Countries criticize each other. Sometimes they are telling the truth and other times they have another motive. In your opinion, what is the percent chance that these countries are telling the truth when they criticize another country? (0-100, order of items randomized)
  - The United States
  - China
  - France

- IOs Telling the Truth International organizations accuse countries of breaking international rules. Sometimes they are telling the truth and other times they have another motive. In your opinion, what is the percent chance that these international organizations are telling the truth when they accuse countries of breaking international rules? (0-100, order of items randomized)
  - The International Criminal Court
  - The World Health Organization
  - The European Union

- Themometer We'd like to get your feelings toward certain countries and international organizations on a "feeling thermometer." A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the country or organization at 50 degrees if you don't feel particularly positively or negatively toward them. How do you feel about following countries or international organizations? (order of items randomized)
  - United States
  - The International Criminal Court
  - Russia
  - Israel

We used the standard set of items for cooperative internationalism. This, too, was measured pretreatment.

- Cooperative internationalism (agree/disagree, 5 point scale, order of items randomized)
  - It is essential for my country to work with other countries to solve problems such as overpopulation, hunger, and pollution.

- It is important for countries to work together to tackle global challenges.
- Countries should work together through international organizations.
- Protecting the global environment is very important.
- Helping to improve the standard of living in other countries is very important.

The main manuscript already contains the exact treatment text. Post-treatment, we measured the respondent's posterior beliefs about Russian guilt and the accuracy of information sources. We randomized the order of the outcome measures and the order of items within outcome measures, where appropriate. For accuracy, we used the term "biased" to tap into the concept of accuracy, without using the exact same words as the pre-treatment measures.

- Outcome: Russian Guilt How likely is it that Russian leaders have committed war crimes in Ukraine? (100 point scale)

- Outcome: Accuracy of Source Some sources of information are biased and others are not. On a scale of 0-100, with 100 being the most biased, where would you place the following sources of information?

  - The International Criminal Court
  - US government
  - The media

After that, respondents answered two manipulation check questions and then demographic questions.

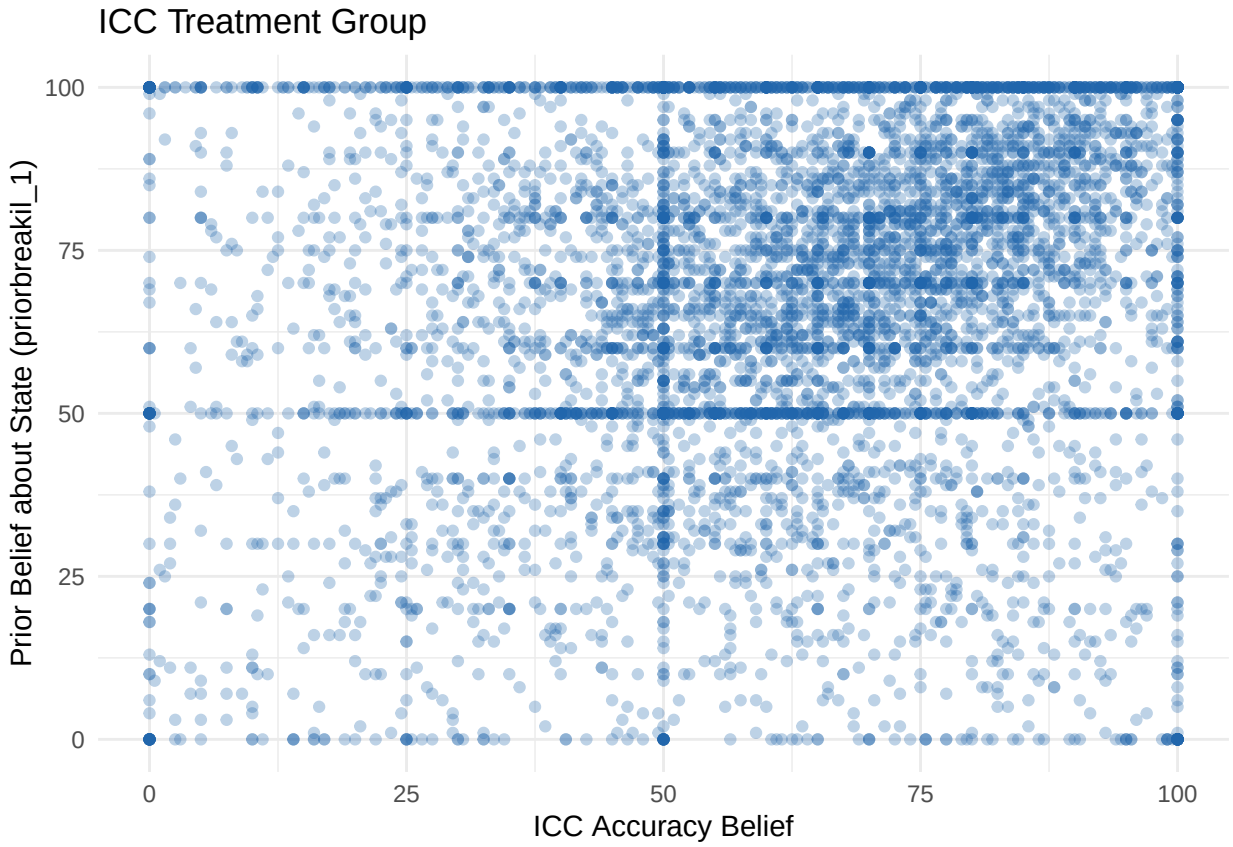
- Manipulation Checks

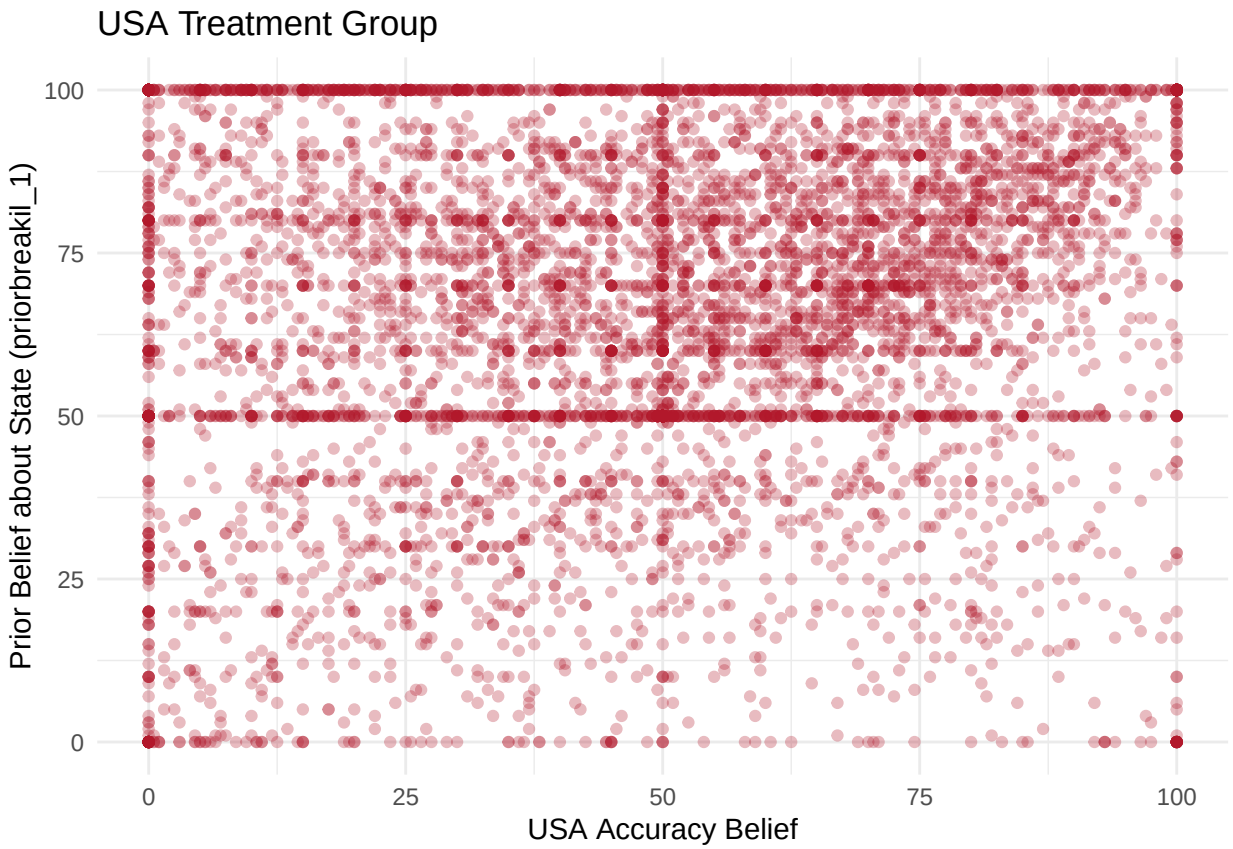
  - In one of the earlier questions, we asked about war crimes. Which country's leaders were accused of committing war crimes in that question? (Russia, USA, Guatemala)
  - In that same earlier question, who was accusing Russian leaders of war crimes? (The International Criminal Court, The US government, The Ukranian government)

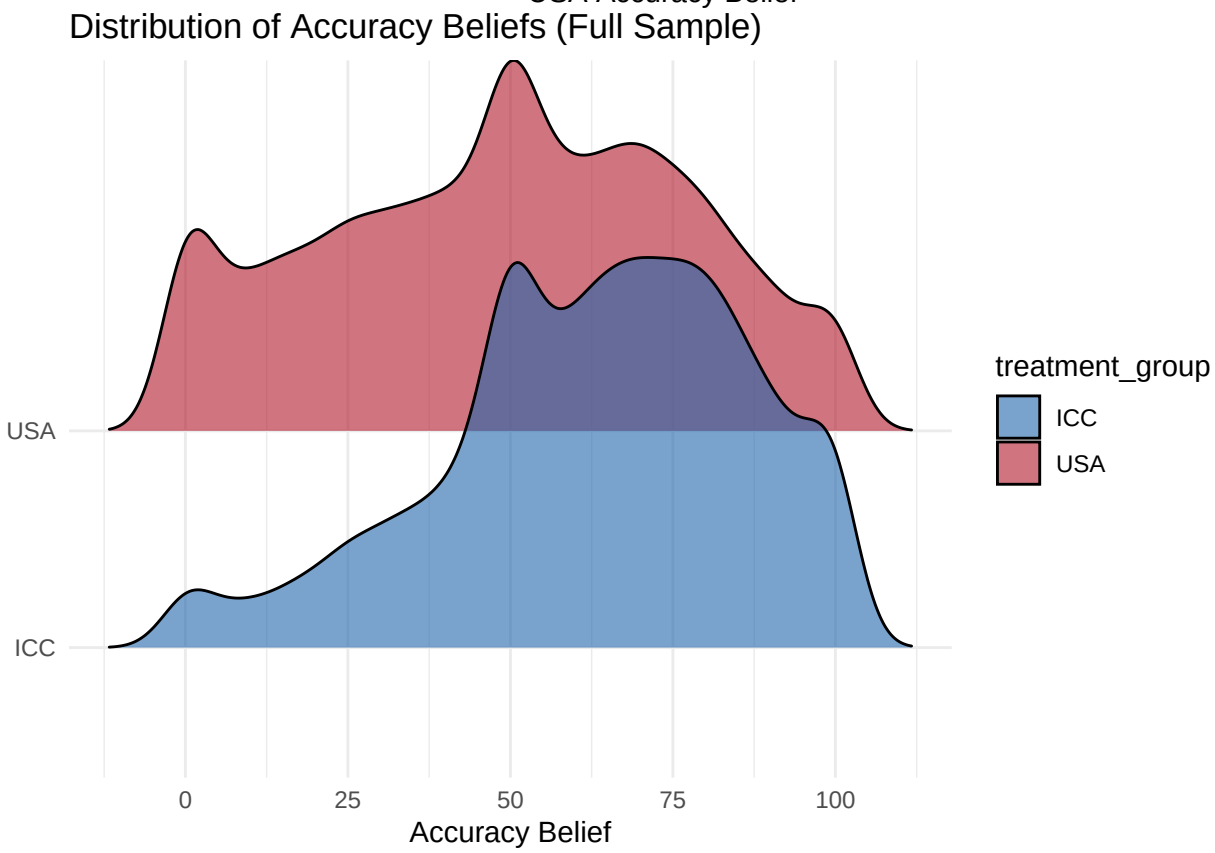
- Demographics
  - Which political party do you feel most closely represents your views? (The lists varied by country.)
  - What is the highest level of education you have completed? (9 point scale, ranging from "No formal schooling" to "Post-graduate")
  - What is your current working status? (6 standard options)
  - What is your approximate monthly income? (12 point scale, currency and ranges varied by country)
  - In political matters, people talk of "the left" and "the right." Please tell me where would you place your views on a 10-point scale where 1 is the 'left' and 10 is the 'right'? (10 point scale)

C Summary statistics

This figure shows the distribution of responses to the pre-treatment questions about prior Russian guilt and perceptions of the two sources. The two things are positively correlated, but not completely so.




USA Treatment Group
Prior Belief about State (priorbreakil_1)
100
75
50
25
0
0
25
50
75
100
USA Accuracy Belief


Distribution of Accuracy Beliefs (Full Sample)
USA
ICC
0
25
50
75
100
Accuracy Belief
treatment_group
ICC
USA

C.1 Country specific summary information

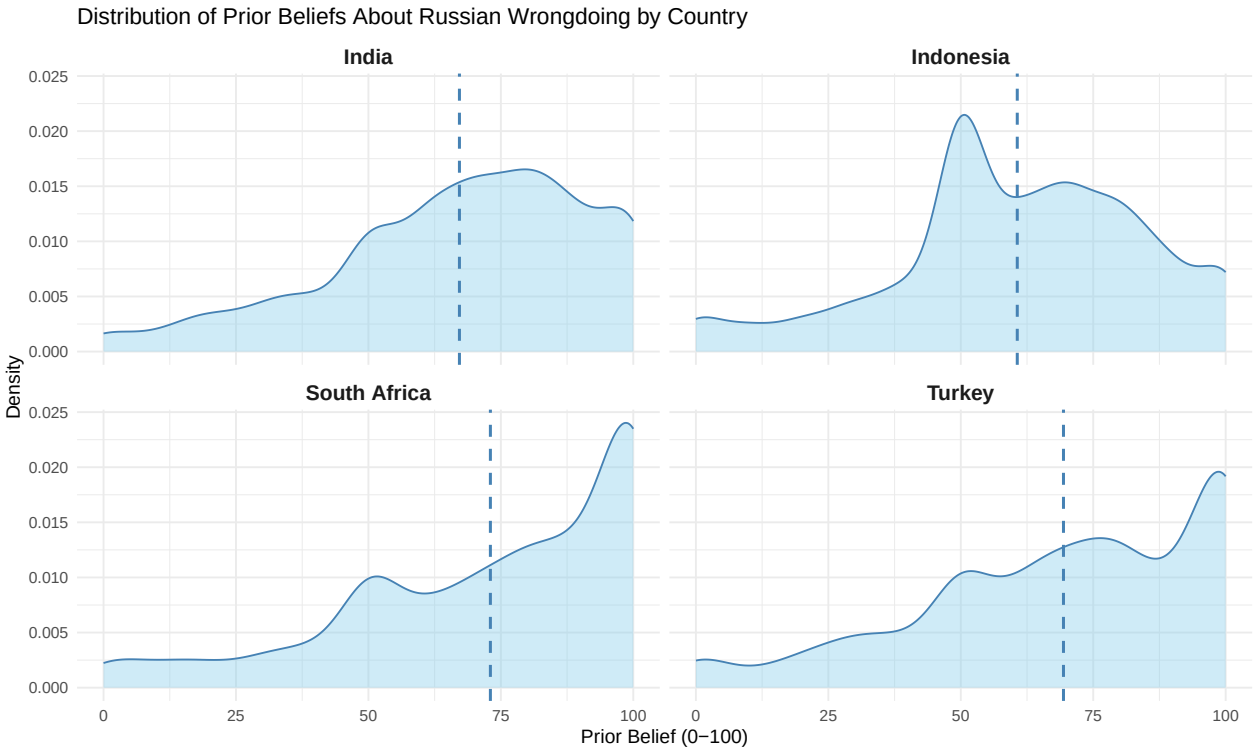


Figure C.1: Distribution of priors about Russian guilt by country.

C.2 Hypocrisy - Trust

Since our pre-treatment items also asked whether the respondents thought the United States had broken international law, we also looked at whether this was correlated with perceptions of trustworthiness of the United States. They are correlated.

D Balance tests

These are balance tests using the approach in Hansen and Bowers (2008). Samples are generally well-balanced in key covariates, while there are some imbalances in individual covariates. For example, there were more women in the ICC treatment group in Turkey, compared to the control group. In Indonesia, respondents in the USA treatment group had slightly higher incomes than the control group.

Table C.1: Beliefs About U.S. Violation of International Law and Trust in the U.S.

	Dependent variable:	
	Trust in the United States	
	No Controls	With Controls
	(1)	(2)
U.S. Violated Intl. Law	-0.212^{***}	-0.171^{***}
	(0.013)	(0.014)
Age		-0.063^{*}
		(0.035)
Female		5.630^{***}
		(0.774)
Education		-0.830^{***}
		(0.196)
Income		0.140
		(0.151)
Voted for Incumbent		13.344^{***}
		(0.810)
Constant	64.417^{***}	63.632^{***}
	(0.968)	(2.227)
Observations	6,553	5,415
R^2	0.038	0.090
Adjusted R^2	0.038	0.089
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table D.1: ICC Treatment Balance Test

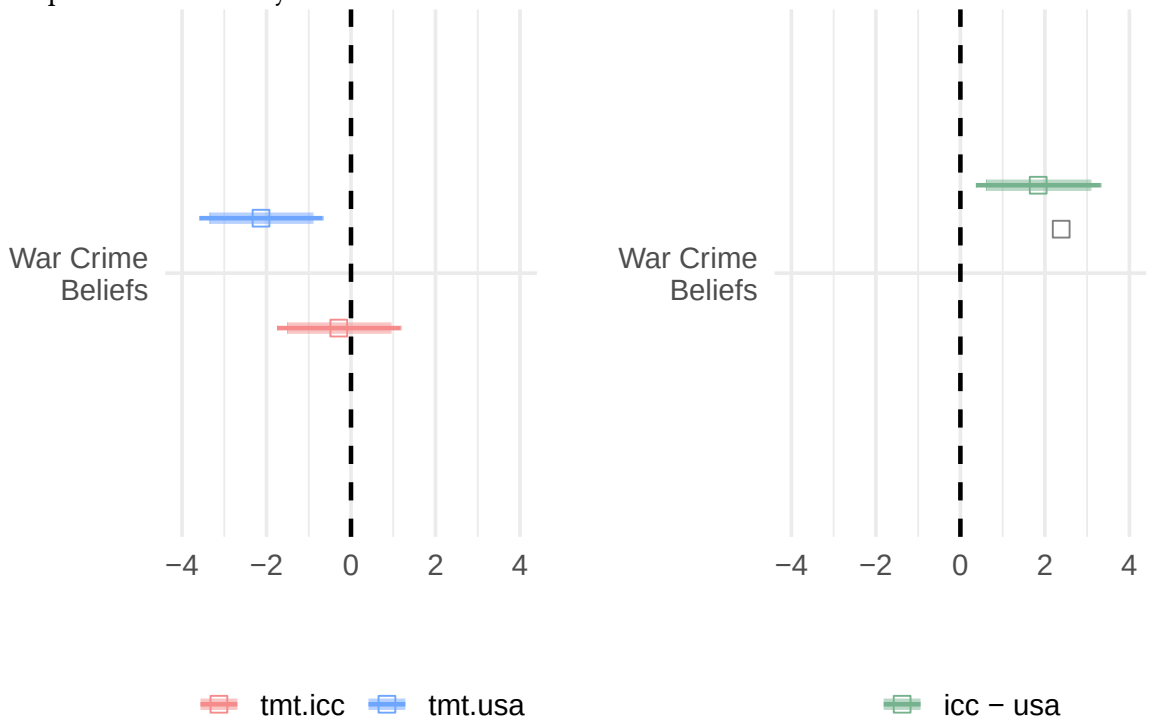
country	covariate	Control	Treated	Std.Diff	Z	p.value
Indonesia	age	36.748	36.102	-0.060	-0.860	0.390
	education	8.222	8.367	0.060	0.863	0.388
	female	0.486	0.512	0.051	0.733	0.464
	income_numeric	6.057	6.415	0.111	1.601	0.109
	incumbent	0.187	0.206	0.048	0.689	0.491
South Africa	age	36.327	36.075	-0.023	-0.355	0.723
	education	5.681	5.617	-0.043	-0.679	0.497
	female	0.544	0.511	-0.065	-1.014	0.310
	income_numeric	6.744	6.634	-0.054	-0.850	0.395
	incumbent	0.325	0.312	-0.029	-0.444	0.657
India	age	35.735	35.583	-0.014	-0.222	0.824
	education	7.450	7.333	-0.082	-1.314	0.189
	female	0.466	0.481	0.031	0.500	0.617
	income_numeric	4.136	3.977	-0.063	-1.017	0.309
	incumbent	0.647	0.600	-0.096	-1.546	0.122
Turkey	age	38.437	38.106	-0.030	-0.427	0.669
	education	8.518	8.426	-0.078	-1.168	0.243
	female	0.445	0.527	0.165	2.335	0.020
	income_numeric	5.201	5.150	-0.033	-0.490	0.624
	incumbent	0.239	0.222	-0.041	-0.573	0.567

Table D.2: USA Treatment Balance Test

country	covariate	Control	Treated	Std.Diff	Z	p.value
Indonesia	age	36.748	36.649	-0.009	-0.133	0.894
	education	8.222	8.278	0.023	0.326	0.745
	female	0.486	0.452	-0.068	-0.973	0.330
	income_numeric	6.057	6.644	0.186	2.670	0.008
	incumbent	0.187	0.172	-0.039	-0.557	0.578
South Africa	age	36.327	36.021	-0.027	-0.424	0.671
	education	5.681	5.716	0.024	0.379	0.705
	female	0.544	0.514	-0.061	-0.947	0.344
	income_numeric	6.744	6.737	-0.004	-0.057	0.954
	incumbent	0.325	0.303	-0.049	-0.752	0.452
India	age	35.735	35.890	0.014	0.220	0.826
	education	7.450	7.419	-0.023	-0.369	0.712
	female	0.466	0.478	0.025	0.399	0.690
	income_numeric	4.136	4.156	0.008	0.121	0.904
	incumbent	0.647	0.675	0.059	0.945	0.344
Turkey	age	38.437	38.631	0.018	0.250	0.802
	education	8.518	8.304	-0.173	-2.509	0.012
	female	0.445	0.508	0.126	1.773	0.076
	income_numeric	5.201	5.118	-0.053	-0.780	0.435
	incumbent	0.239	0.231	-0.018	-0.251	0.802

E Country-specific intercepts

The main manuscript showed results with a single intercept. These are results with a country-specific intercept. Results are very similar.



F Table of Regression Results from Main Figures

Table F.2 shows the regression results when we regress posteriors about Russian guilt on the ICC and USA treatments together. In other words, these regressions compare the two treatment groups with the control group. Table F.4 shows the same thing, only with support for the policy responses as the outcome measures. Table F.6 and Table F.8 do the same thing, only they exclude control group respondents. In other words, they compare outcomes between the ICC and USA treatment groups only.

Table F.1: Effect of Treatment on War Crimes Beliefs

	Dependent variable:	
	Russia Committed War Crimes	
	No Controls	With Controls
	(1)	(2)
ICC Treatment	−0.282	−0.810
	(0.788)	(0.843)
USA Treatment	−2.131***	−2.367***
	(0.790)	(0.846)
Age		0.031
		(0.032)
Female		7.669***
		(0.691)
Education		−1.435***
		(0.175)
Income		0.191
		(0.135)
Incumbent Voted For		0.801
		(0.723)
Constant	68.722***	74.241***
	(0.558)	(1.883)
Observations	6,508	5,415
R^2	0.001	0.035
Adjusted R^2	0.001	0.034
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.2: Effect of Treatment on War Crimes Beliefs

	Dependent variable:	
	Russia Committed War Crimes	
	No Controls	With Controls
	(1)	(2)
ICC Treatment	−0.282	−0.810
	(0.788)	(0.843)
USA Treatment	−2.131***	−2.367***
	(0.790)	(0.846)
Age		0.031
		(0.032)
Female		7.669***
		(0.691)
Education		−1.435***
		(0.175)
Income		0.191
		(0.135)
Incumbent Voted For		0.801
		(0.723)
Constant	68.722***	74.241***
	(0.558)	(1.883)
Observations	6,508	5,415
R^2	0.001	0.035
Adjusted R^2	0.001	0.034
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.3: Effect of Treatment on Policy Preferences

	Dependent variable:					
	Non-mil. Aid	Non-mil. Aid	Mil. Aid	Mil. Aid	Sanctions	Sanctions
	(1)	(2)	(3)	(4)	(5)	(6)
ICC Treatment	0.068*	0.057	0.019	0.052	0.063	0.013
	(0.039)	(0.040)	(0.039)	(0.042)	(0.044)	(0.043)
USA Treatment	−0.025	−0.024	−0.022	−0.067	−0.039	−0.042
	(0.039)	(0.040)	(0.039)	(0.043)	(0.044)	(0.043)
Age				0.010***	−0.009***	−0.001
				(0.002)	(0.002)	(0.002)
Female				0.016	0.347***	0.349***
				(0.035)	(0.036)	(0.035)
Education				0.039***	−0.024***	0.004
				(0.009)	(0.009)	(0.009)
Income				0.007	−0.002	0.011
				(0.007)	(0.007)	(0.007)
Incumbent Voted For				−0.108***	0.219***	0.006
				(0.036)	(0.037)	(0.037)
Constant	3.344***	2.882***	3.012***	2.745***	3.190***	2.824***
	(0.027)	(0.028)	(0.028)	(0.095)	(0.097)	(0.096)
Observations	6,516	6,516	6,516	5,415	5,415	5,415
R^2	0.001	0.001	0.0002	0.016	0.033	0.019
Adjusted R^2	0.001	0.0004	−0.0001	0.014	0.032	0.018

Note: *p<0.1; **p<0.05; ***p<0.01

Table F.4: Effect of Treatment on Policy Preferences

	Dependent variable:					
	Non-mil. Aid	Non-mil. Aid	Mil. Aid	Mil. Aid	Sanctions	Sanctions
	(1)	(2)	(3)	(4)	(5)	(6)
ICC Treatment	0.068*	0.057	0.019	0.052	0.063	0.013
	(0.039)	(0.040)	(0.039)	(0.042)	(0.044)	(0.043)
USA Treatment	−0.025	−0.024	−0.022	−0.067	−0.039	−0.042
	(0.039)	(0.040)	(0.039)	(0.043)	(0.044)	(0.043)
Age				0.010***	−0.009***	−0.001
				(0.002)	(0.002)	(0.002)
Female				0.016	0.347***	0.349***
				(0.035)	(0.036)	(0.035)
Education				0.039***	−0.024***	0.004
				(0.009)	(0.009)	(0.009)
Income				0.007	−0.002	0.011
				(0.007)	(0.007)	(0.007)
Incumbent Voted For				−0.108***	0.219***	0.006
				(0.036)	(0.037)	(0.037)
Constant	3.344***	2.882***	3.012***	2.745***	3.190***	2.824***
	(0.027)	(0.028)	(0.028)	(0.095)	(0.097)	(0.096)
Observations	6,516	6,516	6,516	5,415	5,415	5,415
R^2	0.001	0.001	0.0002	0.016	0.033	0.019
Adjusted R^2	0.001	0.0004	−0.0001	0.014	0.032	0.018

Note: *p<0.1; **p<0.05; ***p<0.01

Table F.5: Effect of ICC Treatment on War Crimes Beliefs (Restricted to tmt.ctr = 0)

	Dependent variable:	
	Russia Committed War Crimes	
	No Controls	With Controls
	(1)	(2)
ICC Treatment	1.849**	1.548*
	(0.792)	(0.851)
Age		0.046
		(0.039)
Female		7.431***
		(0.853)
Education		−1.316***
		(0.216)
Income		0.084
		(0.165)
Incumbent Voted For		0.554
		(0.894)
Constant	66.591***	71.251***
	(0.561)	(2.272)
Observations	4,344	3,614
R^2	0.001	0.031
Adjusted R^2	0.001	0.029
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.6: Effect of ICC Treatment on War Crimes Beliefs (Restricted to tmt.ctr = 0)

	Dependent variable:	
	Russia Committed War Crimes	
	No Controls	With Controls
	(1)	(2)
ICC Treatment	1.849**	1.548*
	(0.792)	(0.851)
Age		0.046
		(0.039)
Female		7.431***
		(0.853)
Education		−1.316***
		(0.216)
Income		0.084
		(0.165)
Incumbent Voted For		0.554
		(0.894)
Constant	66.591***	71.251***
	(0.561)	(2.272)
Observations	4,344	3,614
R^2	0.001	0.031
Adjusted R^2	0.001	0.029
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table F.7: Effect of ICC Treatment on Policy Preferences (Restricted to tmt.ctr = 0)

	Dependent variable:					
	Non-mil. Aid	Non-mil. Aid	Mil. Aid	Mil. Aid	Sanctions	Sanctions
	(1)	(2)	(3)	(4)	(5)	(6)
ICC Treatment	0.093**	0.082**	0.040	0.119***	0.103**	0.055
	(0.038)	(0.040)	(0.039)	(0.042)	(0.044)	(0.043)
Age				0.011***	−0.008***	−0.0004
				(0.002)	(0.002)	(0.002)
Female				0.009	0.322***	0.341***
				(0.042)	(0.044)	(0.043)
Education				0.053***	−0.018	0.003
				(0.011)	(0.011)	(0.011)
Income				0.003	−0.006	0.009
				(0.008)	(0.008)	(0.008)
Incumbent Voted For				−0.094**	0.228***	−0.014
				(0.044)	(0.046)	(0.046)
Constant	3.319***	2.858***	2.990***	2.555***	3.116***	2.784***
	(0.027)	(0.028)	(0.028)	(0.112)	(0.116)	(0.116)
Observations	4,349	4,349	4,349	3,614	3,614	3,614
R^2	0.001	0.001	0.0002	0.021	0.031	0.018
Adjusted R^2	0.001	0.001	0.00001	0.020	0.030	0.016

Note: *p<0.1; **p<0.05; ***p<0.01

Table F.8: Effect of ICC Treatment on Policy Preferences (Restricted to tmt.ctr = 0)

	Dependent variable:					
	Non-mil. Aid	Non-mil. Aid	Mil. Aid	Mil. Aid	Sanctions	Sanctions
	(1)	(2)	(3)	(4)	(5)	(6)
ICC Treatment	0.093**	0.082**	0.040	0.119***	0.103**	0.055
	(0.038)	(0.040)	(0.039)	(0.042)	(0.044)	(0.043)
Age				0.011***	−0.008***	−0.0004
				(0.002)	(0.002)	(0.002)
Female				0.009	0.322***	0.341***
				(0.042)	(0.044)	(0.043)
Education				0.053***	−0.018	0.003
				(0.011)	(0.011)	(0.011)
Income				0.003	−0.006	0.009
				(0.008)	(0.008)	(0.008)
Incumbent Voted For				−0.094**	0.228***	−0.014
				(0.044)	(0.046)	(0.046)
Constant	3.319***	2.858***	2.990***	2.555***	3.116***	2.784***
	(0.027)	(0.028)	(0.028)	(0.112)	(0.116)	(0.116)
Observations	4,349	4,349	4,349	3,614	3,614	3,614
R^2	0.001	0.001	0.0002	0.021	0.031	0.018
Adjusted R^2	0.001	0.001	0.00001	0.020	0.030	0.016

Note: *p<0.1; **p<0.05; ***p<0.01

G Manipulation Checks

We implemented two manipulation checks to assess whether the treatments had their intended effects. The first check asks, *"Which country's leaders were accused of committing war crimes in that question?"* The second asks, *"In that same earlier question, who was accusing Russian leaders of war crimes?"*. The second question was actually quite hard because the choices were "The Ukrainian government, The International Criminal Court, and The US government." Many respondents chose the Ukrainian government.

Table G.1: Mean Manipulation Check Pass Rates

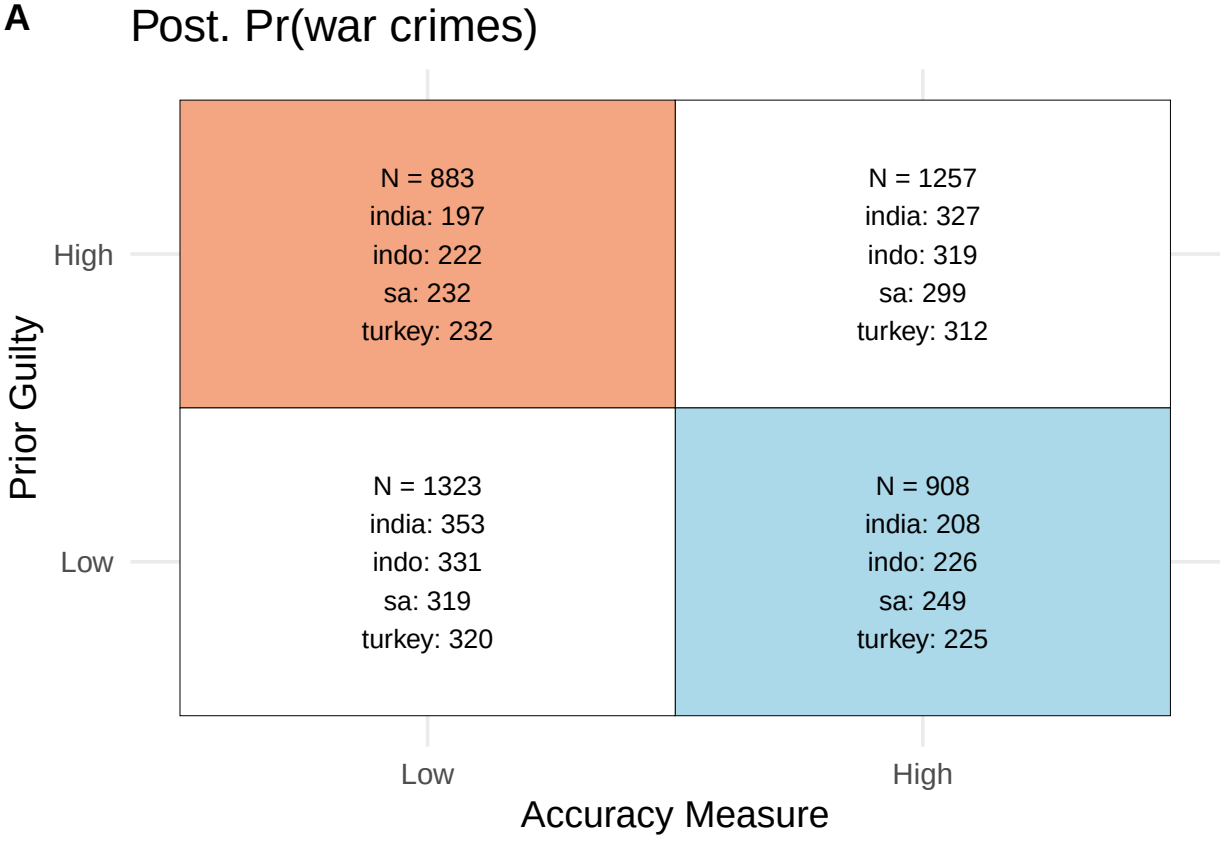
sample	mani_pass1_mean	mani_pass2_mean
Indonesia	0.454	0.553
India	0.750	0.474
Turkey	0.799	0.639
South Africa	0.926	0.614
Overall	0.732	0.570

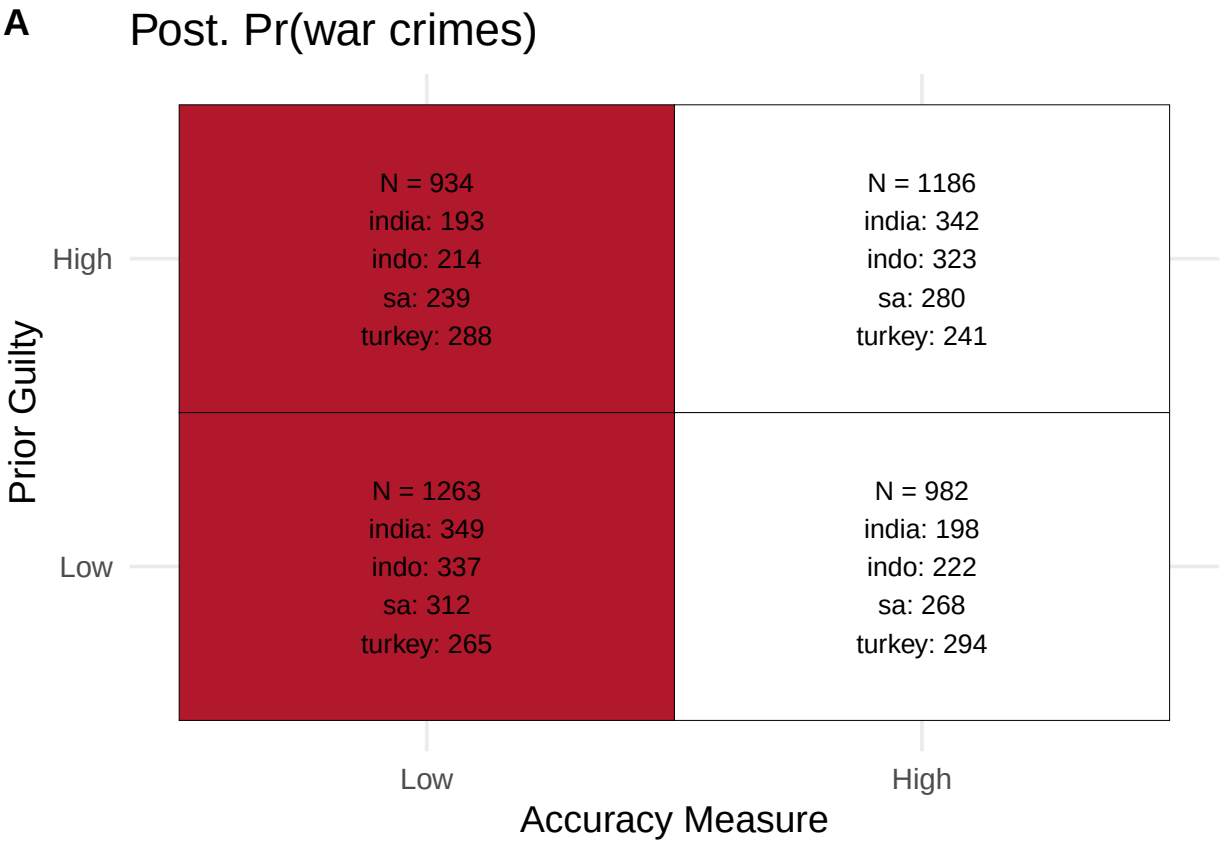
The pass rate for the first and second manipulation check was 73.2% and 57%, respectively.

H Appendix items for H1: Treatment effect on posteriors about guilt and policies

H.1 Sample sizes in the boxes

There are different sample sizes in the four quadrants. The figure below shows the number of observations in each cell, with the same coloring as the first figure. The top pane is for the ICC versus control analysis. The bottom pane is for the USA versus control group analysis.




A
Post. Pr(war crimes)
N = 934
india: 193
indo: 214
sa: 239
turkey: 288
N = 1186
india: 342
indo: 323
sa: 280
turkey: 241
N = 1263
india: 349
indo: 337
sa: 312
turkey: 265
N = 982
india: 198
indo: 222
sa: 268
turkey: 294
High
Low
Prior Guilty
Low
High
Accuracy Measure

H.2 Box plots based on universal medians

In the main manuscript, we classified respondents based on whether they were above/below the medians two prior belief measures, based on country-specific medians. Here, we show the same type of estimates but where respondents are classified based on whether they are above or below global medians. The patterns are similar, though results tend to be a bit weaker.

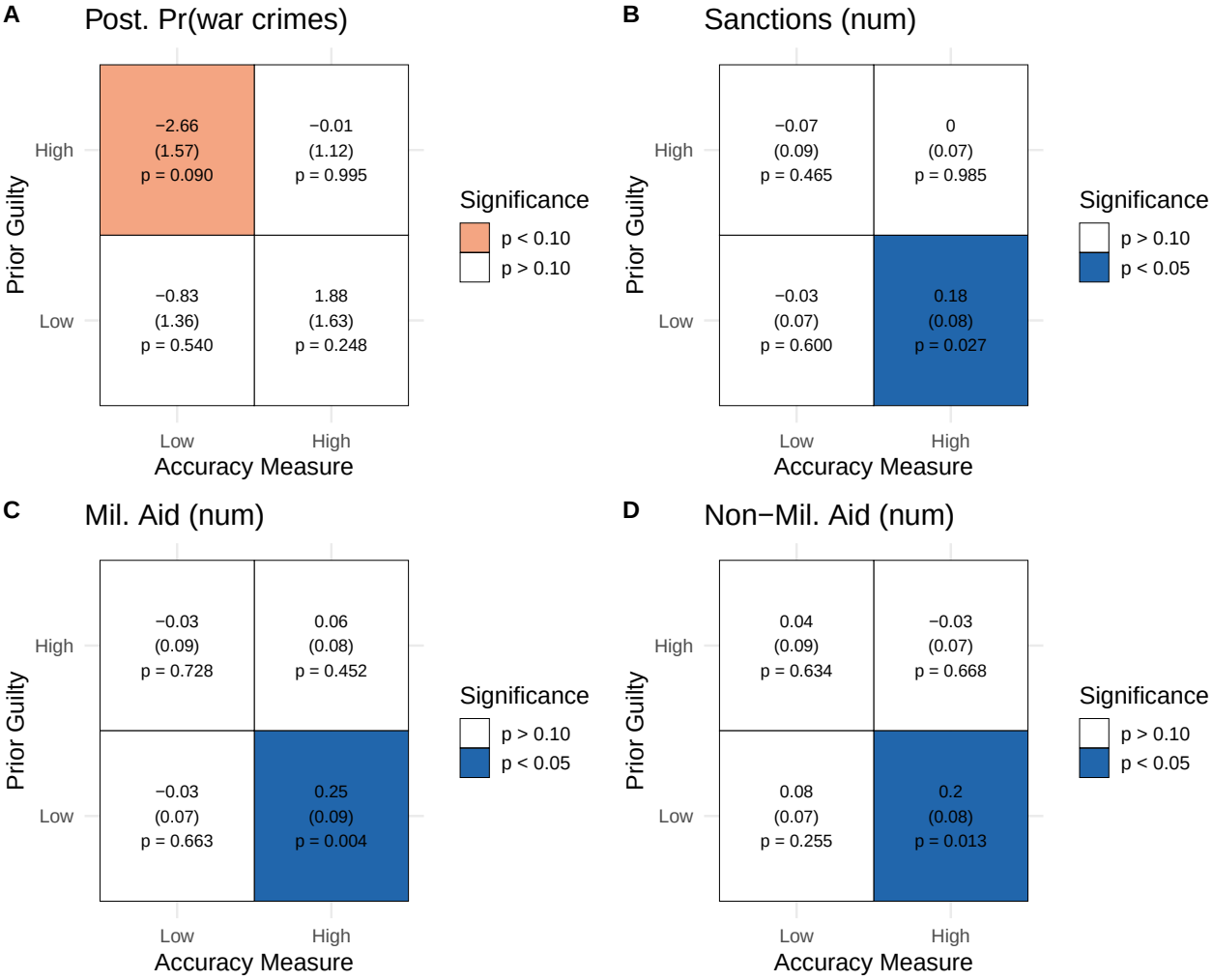


Figure H.1: Effect of ICC treatment, Bayesian boxes (global medians).

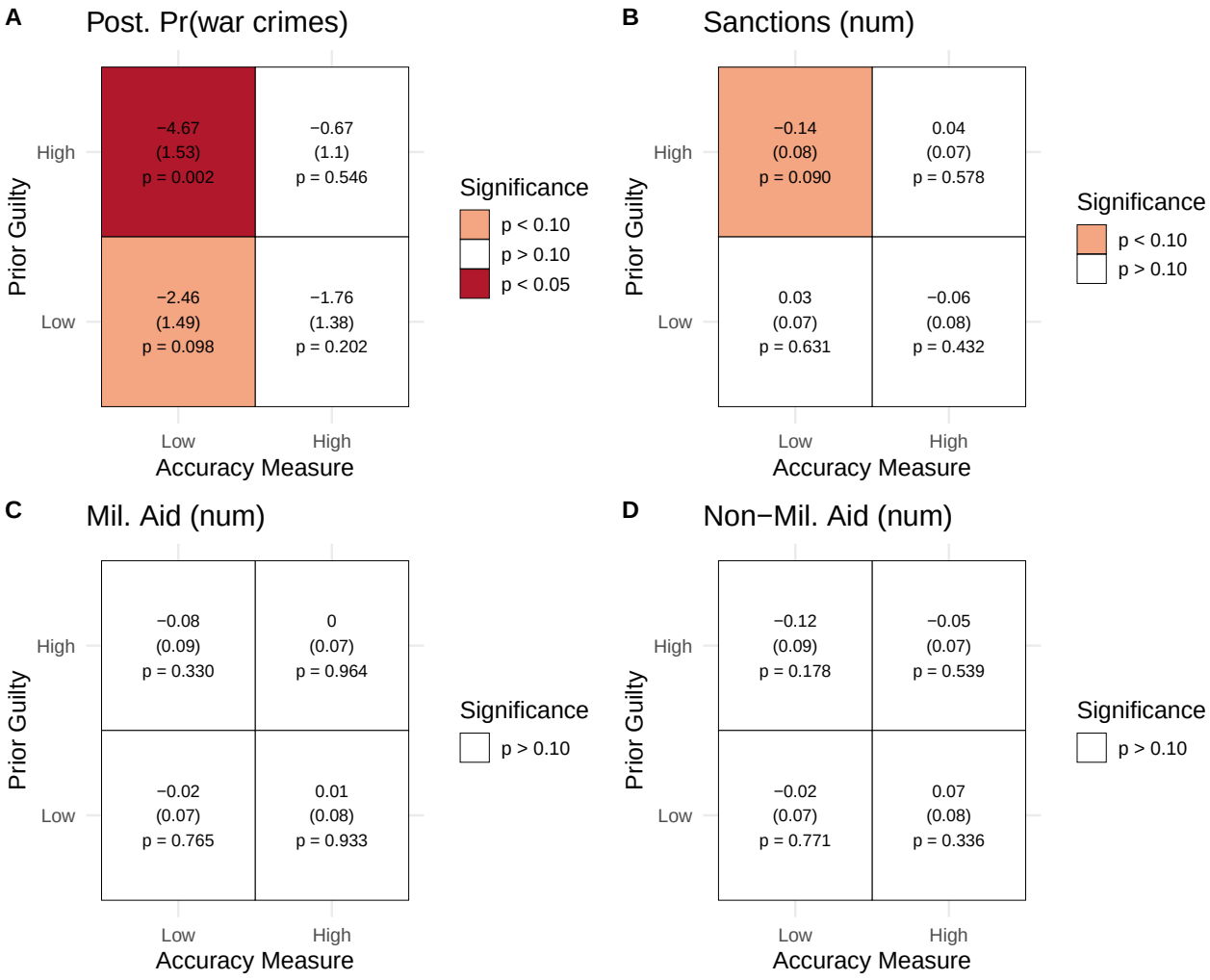


Figure H.2: Effect of USA treatment, Bayesian boxes (global medians).

H.3 Linear interaction terms

Since used linear interaction term models in the Hypothesis 2 analysis, here are the results from those models where the beliefs about Russian guilt is the DV. The lines should be upward sloping and they are.

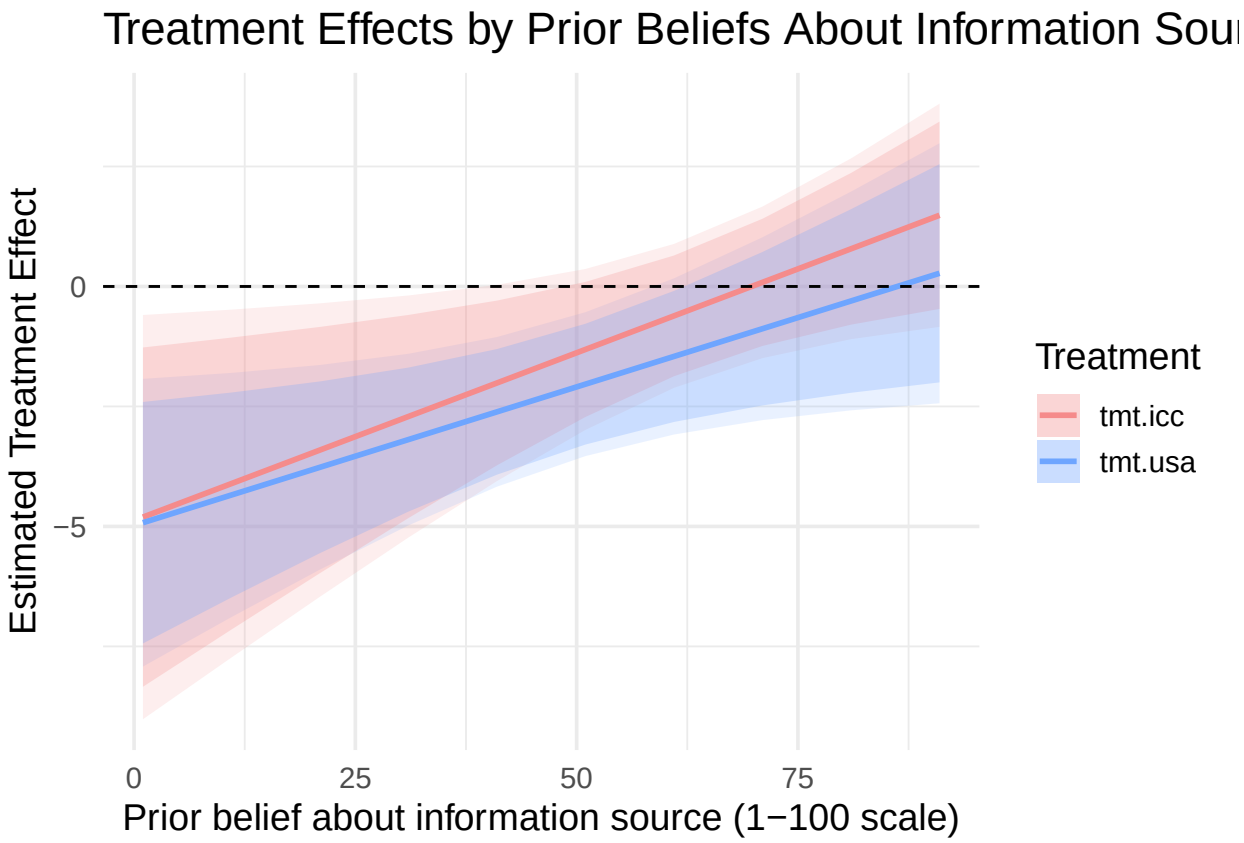
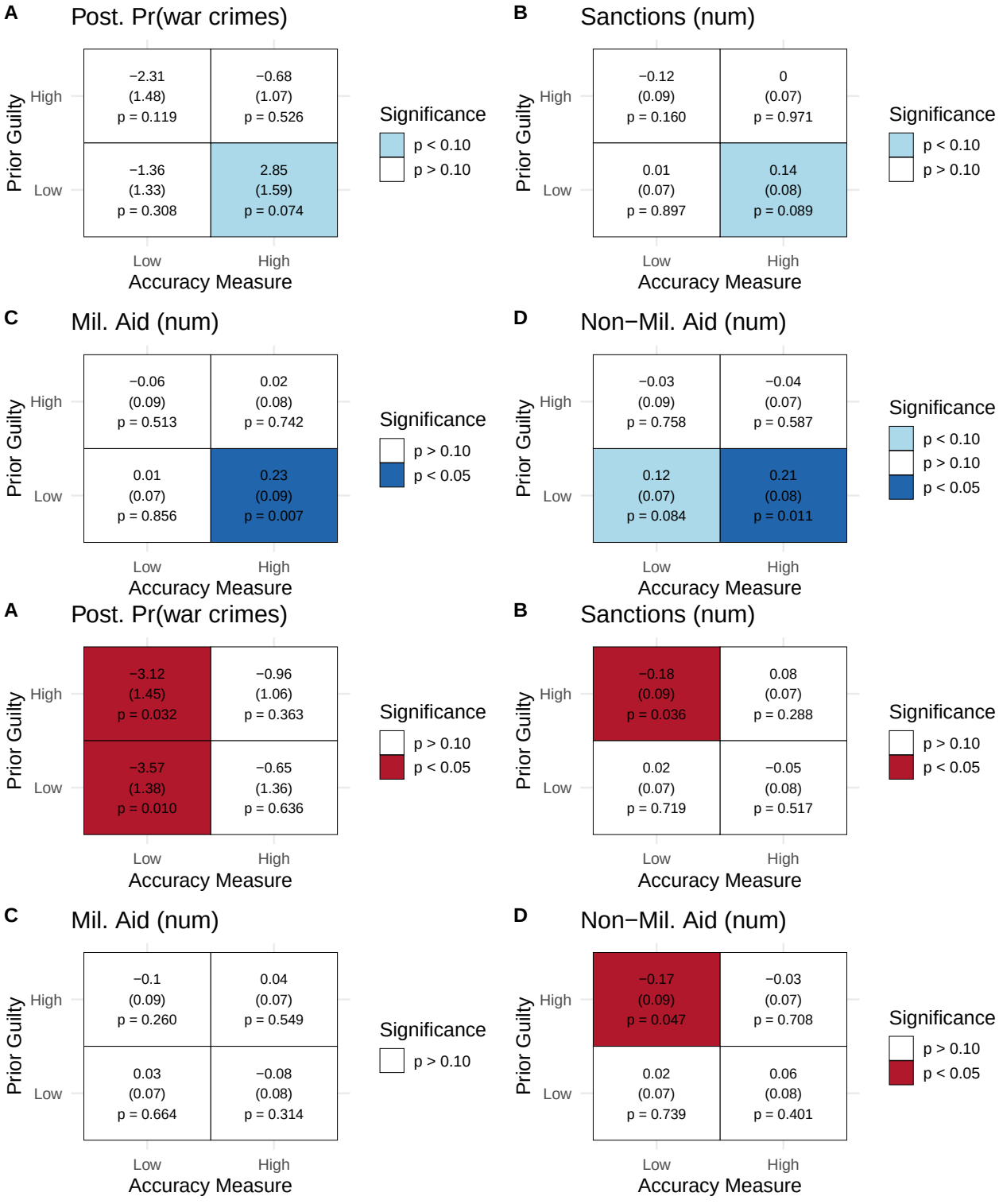


Figure H.3: Effect of treatment on posteriors about Russian guilt, as beliefs about the source vary.

H.4 Box plots with country intercepts in the regressions



I Appendix for H2: Treatment effect on posteriors about the source

I.1 Treatment effects on ICC legitimacy

We also tested whether the ICC treatment influenced perceptions of ICC legitimacy. Figure I.1 shows these estimates graphically. The ICC treatment raised mean legitimacy scores by 0.12 points on a five-point scale (SE = 0.037, p = 0.0013), amounting to roughly a 3.6 percent increase.[57] While the effect size is modest, it is highly significant and consistent with our findings on source trust. When the ICC makes a statement about violations of international law, this meaningfully bolsters public perceptions of its legitimacy.

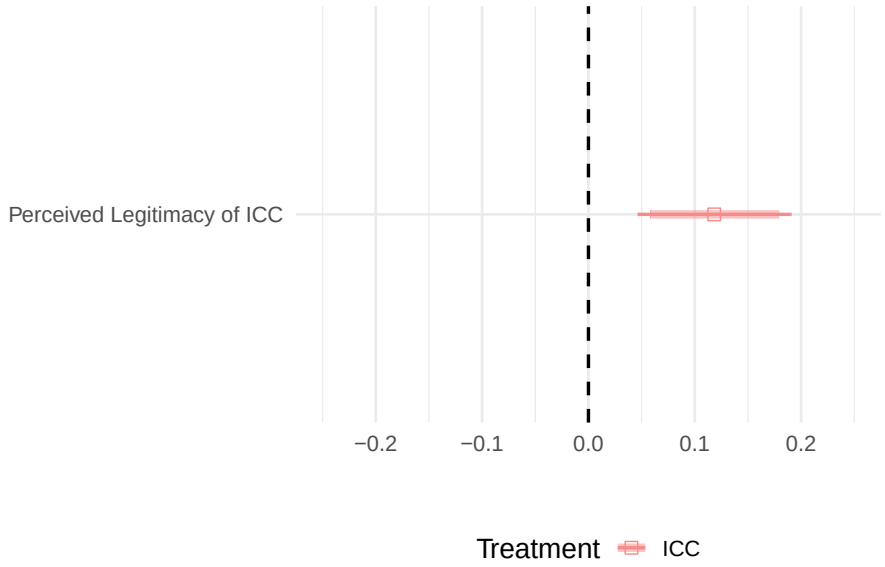


Figure I.1: Effect of treatment on perceptions of ICC legitimacy

I.2 Rolling windows

The main manuscript showed results for Hypothesis 2 using a linear interaction term. Figure I.2 shows estimated treatment effects using a rolling windows approach. We gradually vary the sample used in these regressions, including priors within certain ranges. Figure I.3 shows a linear interaction term regression, where treatment effects vary according to priors about source quality.

[57] We did not conduct the analogous test for U.S. legitimacy.

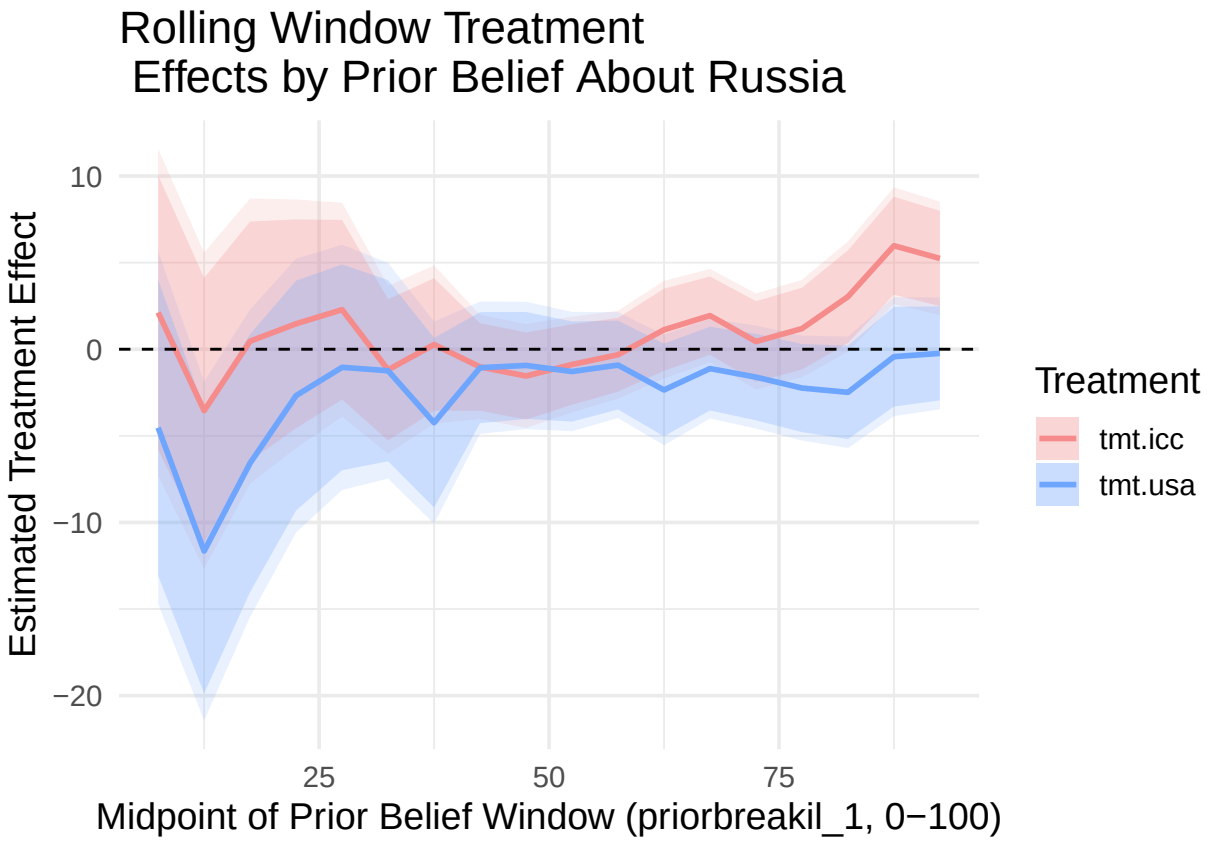


Figure I.2: Effect of treatment on posteriors about source quality, as priors vary.

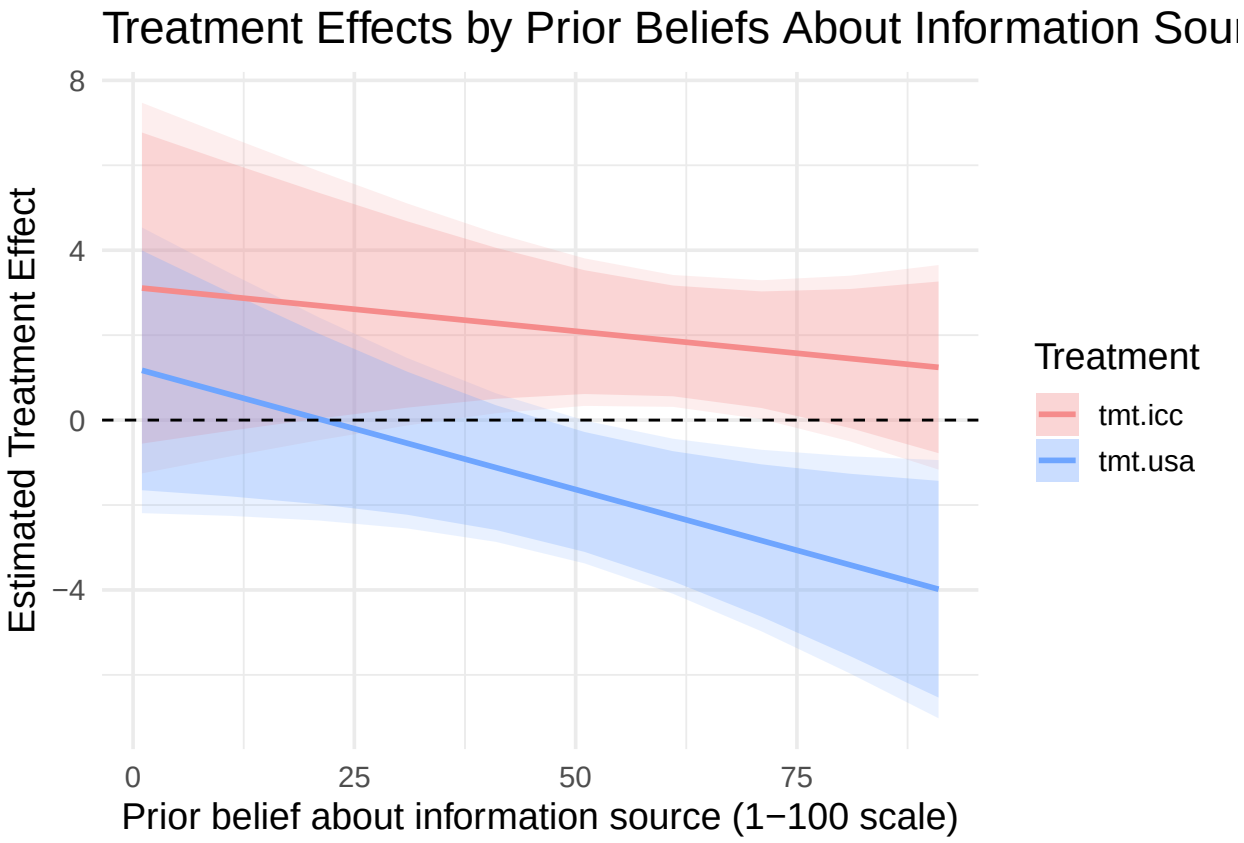


Figure I.3: Effect of treatment on posteriors about source quality, as beliefs about the source vary.

J Appendix for cooperative internationalism moderation

Did cooperative internationalism moderate treatment effects? Our surveys included standard, pre-treatment cooperative internationalism items. We asked about the respondents' agreement (on a 5 point scale) with the statements: (1) "It is essential for my country to work with other countries to solve problems such as overpopulation, hunger, and pollution" (2) "It is important for countries to work together to tackle global challenges," (3) "Countries should work together through international organizations," (4) "Protecting the global environment is very important," and (5) "Helping to improve the standard of living in other countries is very important."[58]

The theoretical model suggests that the effects of CI are ambiguous. One the one hand, respondents that scored higher on the CI measures should show larger treatment effects for the ICC. Presumably, they should have higher pre-treatment beliefs about the accuracy of the ICC's information, which should magnify the ICC treatment effect. This is analogous to the argument most commonly found in existing research. On the other hand, they also are likely to already have higher prior beliefs about Russian guilt, which has a non-monotonic effect on the magnitude of predicted treatment effects. It could mute treatment effects for respondents that already strongly believe in Russian guilt. In our sample, both of these correlations were apparent. Higher CI respondents had higher beliefs about the accuracy of the ICC and higher prior beliefs in Russian guilt.

Note that this same ambiguity applies to moderation based on partisanship.[59] A respondent's party identification could affect their perception of sources. In South Africa, the African National Congress (ANC) is generally less aligned with the United States than the Democratic Alliance (DA). On the one hand, this could mean that ANC members would be less moved by information from the United States. On the other hand, their members may not already have a deeply held belief that Russia is guilty of war crimes, which means their opinions are more movable. A DA member might be more trusting of the United States, which magnifies treatment effects. But they may already think Russia is guilty, muting treatment effects. Increasing trust in a source of information unambiguously increases the treatment effect of information from that source. But moving prior beliefs has a non-monotonic effect on treatment effects. Treatment effects are biggest for people with moderate prior beliefs. Since many moderating variables, like party identification, are correlated with both, their net effect is hard to predict, theoretically.

Figure J.1 shows the estimated treatment effects, broken down by whether respondents were above or below the average score on the CI items. The left pane shows effects on posterior beliefs about Russian war crimes. The right pane shows effects on the policy responses.

Cooperative internationalism has inconsistent moderating effects. Looking first at only the ICC treatment effects, for three of the four outcome measures, higher CI respondents had weaker ICC treatment effects. This is contrary to expectations that are based only on a theory that links CI with perceptions of an IO's credibility. On the other hand, this would be consistent with a theory that links CI to a ceiling effect, where high CI respondents already believe Russia is guilty, so they can't raise this posterior probability much higher.

Looking next at a comparison of ICC and USA treatment effects, the results are also inconsistent in their support or disconfirmation of arguments about CI. On the one hand, the ICC treatment ef-

[58] We randomized the order of these items. We did not ask about militant internationalism, since our focus was on international law.

[59] For examples where partisanship moderates the effects of an informational treatment, see Chaudoin (2023) and Brutger (2021).

fect was generally larger than the USA treatment effect for high CI respondents. However, the ICC effect was larger than the ICC effect among low CI respondents for two out of four outcomes (beliefs about Russian guilt and non-military aid). For those outcome measures, the difference between ICC and USA treatment effects for high CI respondents was comparable to the difference for low CI respondents.[60]

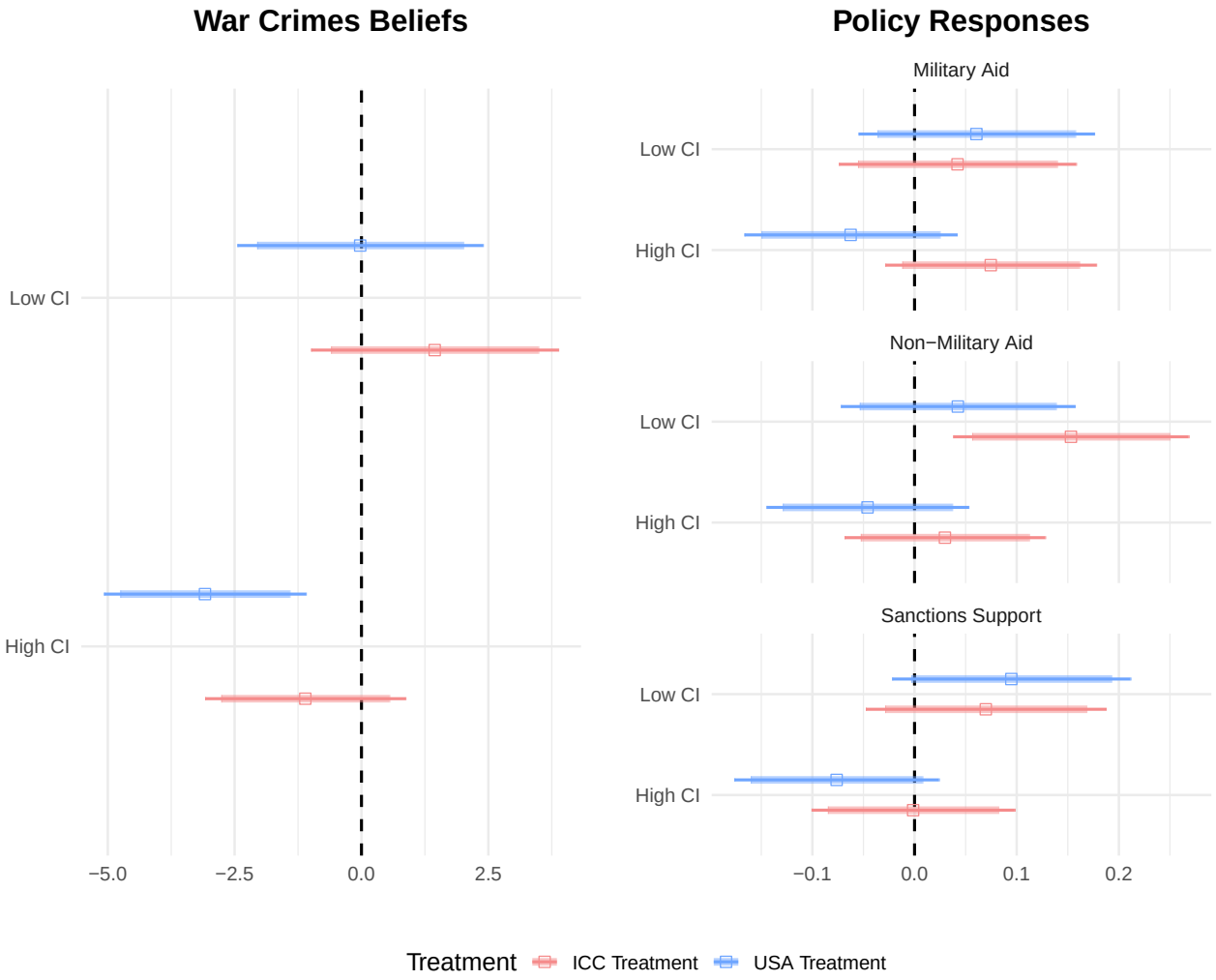
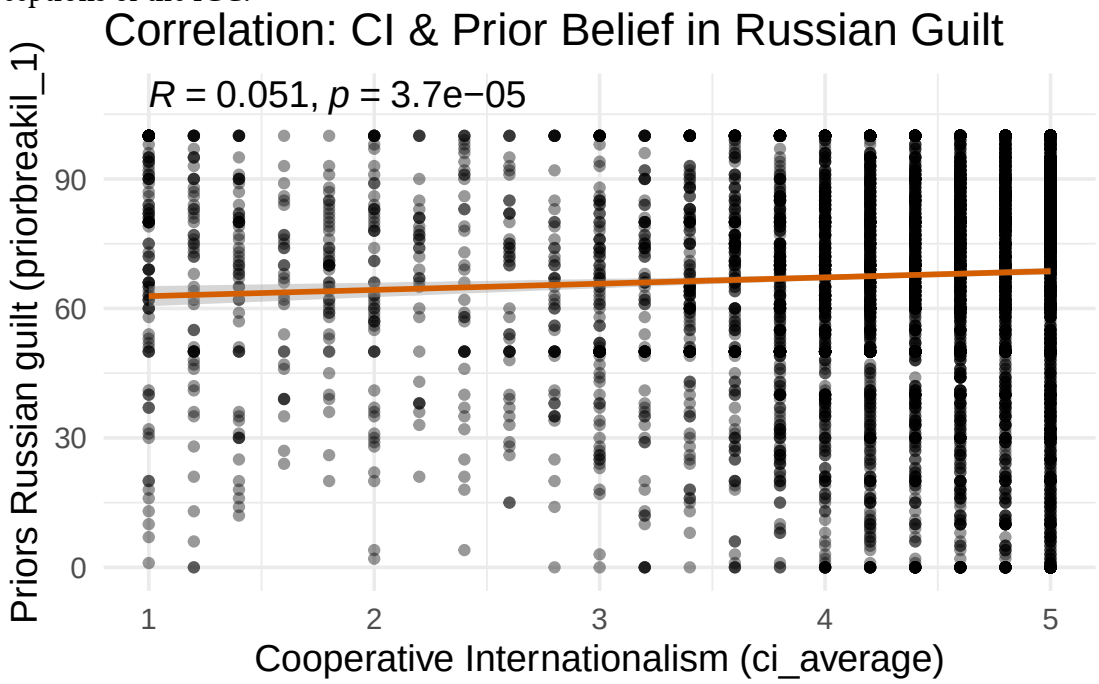


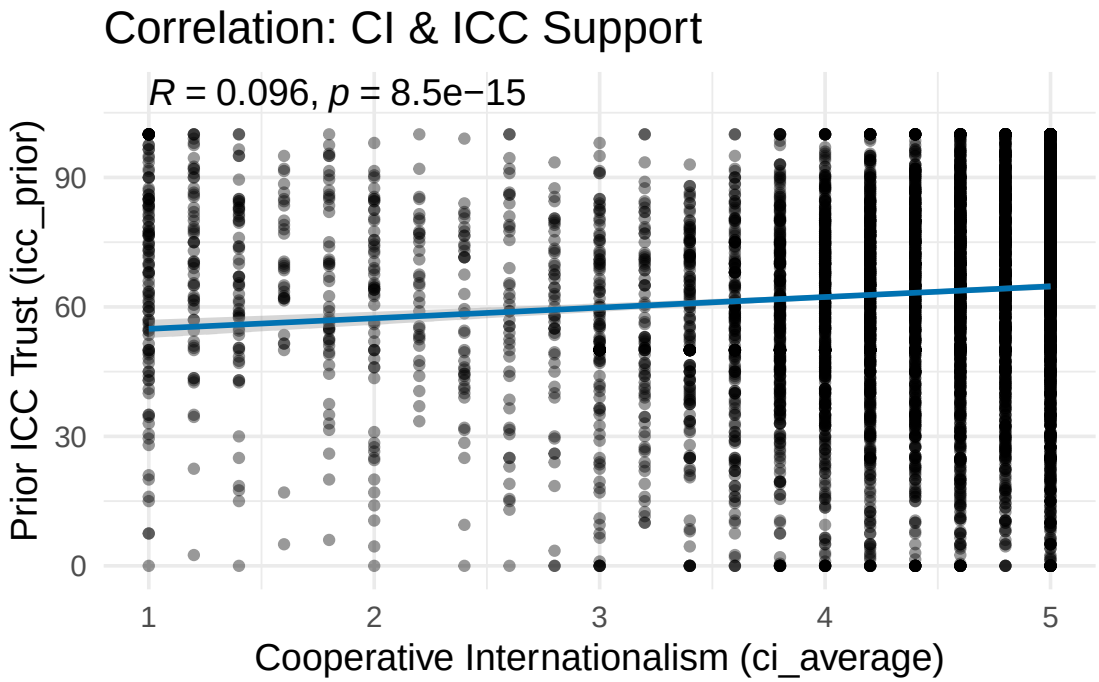
Figure J.1: Treatment effects, broken down by cooperative internationalism

Our point here is not that CI contains no useful information or that it has no effect on attitudes towards foreign policies. On the contrary, it is well-correlated with important parameters, like prior beliefs about the world or about the accuracy of sources. CI is a good predictor of foreign policy attitudes. However, it is theoretically ambiguous as a moderator of treatment effects. CI is a bundle of things related to priors, and therefore its net impact on predicted treatment effects is theoretically ambiguous. CI also likely contains other things that moderate treatment effects in ways that go beyond Bayesian updating. In our application, this ambiguity was born out, even though CI was correlated with prior attitudes as expected.

[60] In the appendix, we also estimate 2x2 boxes using CI and prior beliefs as the two moderating variables. There are not clear patterns.

Above, we stated that correlations between cooperative internationalism and prior beliefs about Russian guilt / perceptions of the ICC were as we would expect. We show that here. Higher CI respondents were more likely to believe that Russia was guilty *ex ante* and they had higher pre-treatment perceptions of the ICC.





We also re-estimated the box plots from the main manuscript, using the CI measure instead of our pre-treatment measures of source accuracy. We do not find the same patterns.

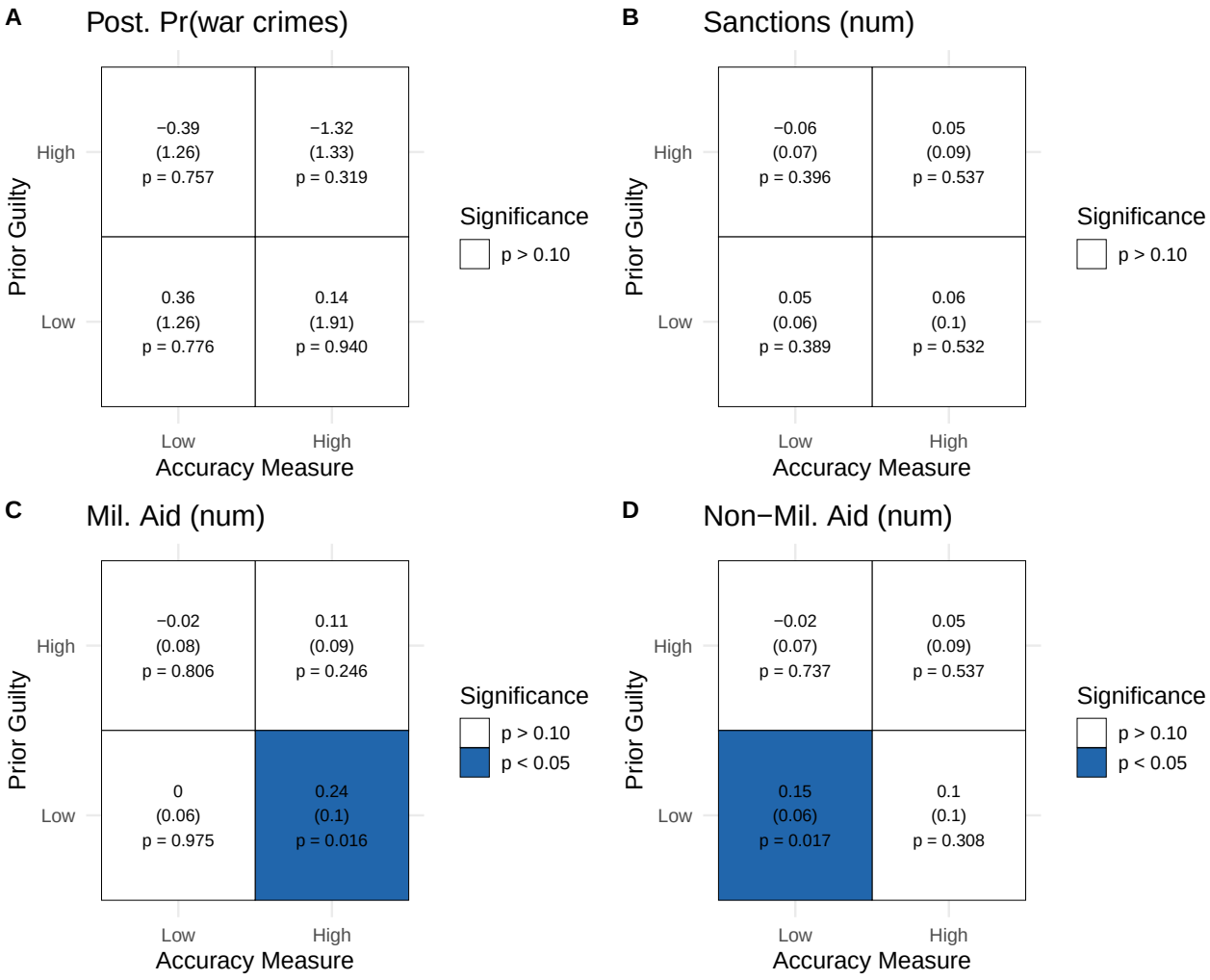


Figure J.2: Effect of ICC treatment, cooperative internationalism boxes.