Leader Influence and
Reputation Formation in World Politics

Jonathan Renshon[1], Allan Dafoe[2], and Paul Huth[3]

[1]Department of Political Science, University of Wisconsin-Madison

[2]Department of Political Science, Yale University

[3]Department of Government & Politics, University of Maryland

Word count: 8840

Accepted at *American Journal of Political Science*

Abstract

The study of reputation is one of the foundational topics of modern international relations. However, fundamental questions remain, including the question of to whom reputations adhere: states, leaders, or both? We offer a theory of *influence-specific reputations* (ISR) that unifies competing accounts of reputation formation. We theorize that reputations will adhere more to actors who are more influential in the relevant decision-making process. We employ two survey experiments, one abstract and one richly detailed involving a US-Iran conflict, to evaluate ISR. We find evidence of large country-specific reputations and moderately sized leader-specific reputations. Consistent with the theory of influence-specific reputations, leader-specific reputations are more important when leaders are more influential.

The data, code, and any additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: http://dx.doi.org/10.7910/DVN/BAG0DS. The authors can be reached at renshon@wisc.edu, allan.dafoe@yale.edu and phuth@umd.edu, respectively. Renshon and Dafoe contributed equally. We are grateful for input from Jonathan Caverley, Guy Grossman, Sophia Hatz, Richard Herrmann, Maria Hjelm, Josh Kertzer, Ryan Powers, Todd Sechser, Rachel Stein, Alastair Smith, Robert Trager, Michael Tomz, Ben Valentino, Steve Weber, Alex Weisiger, Keren Yarhi-Milo, Remco Zwetsloot, and the participants of the Department of Peace and Conflict Research workshop at Uppsala University, the Harvard Security Studies workshop, the UW-Madison International Relations Colloquium, the Survey Experiments Workshop at the 2015 Peace Science Convention, and the Division of Social Science seminar at Hong Kong University of Science & Technology. We thank Stephen Walt for advertising an earlier version of this survey on his Foreign Policy blog.

1 Intro

The study of reputation — beliefs about a trait or tendency of an actor, informed by observation of the actor's past behavior (Dafoe, Renshon, and Huth, 2014, 372) — is one of the foundational topics of modern IR. Despite a recent resurgence in interest in reputations, a foundational question remains unanswered: to whom do reputations adhere, leaders, states, or other actors? As articulated by Jervis (1982, 9), "If one president acts boldly, will other states' leaders draw inferences only about him or will they expect his successors to display similar resolve?. . . On these points we have neither theoretically grounded expectations nor solid evidence." If anything, this understates the importance of the question, as it focuses exclusively on the inferences of leaders, while the growing literature on audience costs and the constraining role of public opinion has demonstrated the import of citizens' judgments about reputational issues. Moreover, two reviews make clear that this question has yet to be adequately addressed (Huth, 1997, 78; Dafoe et al., 2014, 385).

The *agent-specificity* of reputations is critical for both the policies of leaders and the utility of our theories in IR. From a policy perspective, such a question addresses the concerns of elites who wonder whether and how much their country's behavior will benefit or harm their country in the future. If reputations adhere mostly to the leader, then the reputational consequences of a leader's actions will dissipate with their removal, dampening the incentives to build or protect reputation since anytime the leader incurs a "bad" reputation — e.g., by backing down in a dispute — the selectorate can replace her and erase the reputation. By contrast, if reputations adhere mostly to the country, then the reputational consequences of a country's actions will echo long after the removal of the leader.

How agents — whether leaders or citizens — make inferences about reputation is also critical to a number of theories of IR. Recent debates about how to find evidence of audience costs may be misguided, not just because of strategic selection (Schultz, 2001), but because those theories assume that reputations adhere entirely to leaders. Our results strongly suggest the need to account for situations when reputations stick to the country and the leader differentially. Our puzzle — in effect, how reputations are affected by a "break with the past" — is also critical to resolving the well-known theoretical puzzle of whether reputations "matter," since it gives scholars insight as to where to look for the effects of reputation, as well as where they are likely to find null results even when reputations do, in fact, matter.

A brief survey of diplomatic history suggests that reputational inferences adhere to both leaders and states: leaders puzzle over how to interpret the actions of new leaders in light of previous leaders, while new leaders attempt to manipulate estimates of resolve by emphasizing continuity or change from previous leaders. We find that when the USSR launched a campaign to discredit Stalin following his death, U.S. officials wondered whether the reason was "the hope of Soviet leaders to gain respectability abroad by virtue of a complete break with the past" (Office of the Historian, United States Department of State, 1956). A survey of de-classified U.S. diplomatic records finds U.S. leaders struggling to estimate the balance between the impact of Stalin and Khrushchev's forceful personalities and the strong system of Communist party leadership in which they were enmeshed, as well as how those factors shifted over time as the Soviet political system gradually opened up.[1]

Reputations may adhere to any kind of agent: a leader, a ruling group, an organization, a clan, an army, a country, a people, institutions, or any other factor that shapes a country's behavior, such as organizational cultures. The vast majority of work on reputations in international relations has focused on what we term *country-specific reputations* (CSR): viewing reputations as (approximately) adhering to the country as a whole. Some more recent work (e.g., Guisinger and Smith, 2002; McGillivray and Smith, 2008; Wolford, 2007) has theorized reputation as adhering to individual leaders, what we term *leader-specific reputations* (LSR).

While previous research has largely divided itself into the opposing CSR and LSR camps,

[1] E.g., in describing the influence of Khrushchev: "[he is] not quite comparable to Mr. Stalin and did not have the same measure of arbitrary control that the latter had." See Office of the Historian, United States Department of State (1960).

there is both little middle ground, and no general theory of *when* reputations are likely to adhere to any particular entity; i.e., when are reputations more likely to be state- or leader-specific? And over thirty years after the question was posed by Jervis (1982, 9), we still lack empirical research that provides a systematic evaluation of the agent-specificity of reputation. We address these lacuna by introducing a general theory of *Influence Specific Reputation* (ISR), formally deriving testable hypotheses, and examining these hypotheses in an experimental setting.

Our theory posits that reputations attach most to those actors most influential in the relevant decision-making process. Underpinning this is the logic that inferences are most useful when they are accurate and most accurate when they correspond most closely to the decision-making process in question. In regimes where a single leader exerts dominant control, it would be sensible for observers to draw leader-specific reputational inferences; in cases where the leader's influence is constrained by institutions or long-lived groups, country-specific inferences would be more accurate and useful.[2] This analytic distinction extends to issue areas as well; in domains where leaders of democracies tend to be relatively less constrained, such as national security, reputational inferences are likely to be more leader-specific than in other domains, such as monetary policy. Our theory is rationalist at heart, and makes no assumptions about *whose* inferences matter. Our predictions are identical for citizens and leaders alike, with the caveat that if elites do, indeed, adhere more closely to rationalist ideals when the stakes are high (Press, 2005, 158), the results we find for samples of the general population are apt to be under-estimates of their true effect among world leaders.

We use two scenario-based survey experiments to demonstrate the existence of country

[2] A second important factor is the extent to which leader traits, worldview, and strategies are correlated over time. In countries where they are highly correlated, such as may arise if leaders come from the same educational system or cultural group, then we should expect more country-specific reputation. In countries where these are less correlated (and leader influence is high), we should expect more leader-specific reputation. This paper's empirical strategy seeks to hold "all else equal", including this feature; our scenario design should hold correlation of leader type constant across treatment levels. Future work could explore how correlation of leader type might be shaped.

and leader reputations, as well as tease apart their relative importance. The experiments provide a direct test of our theory by evaluating whether country- and leader-specific reputations vary in importance depending on the decision-making influence of the leader. The first experiment is an abstract scenario about a conflict between countries **A** and **B**, while the second study represents a more unusual and difficult test for our theory: a realistic and detailed scenario about a conflict between the U.S. and Iran, based on a Brookings Institute wargame. We find nearly identical results across these two designs and samples.

In both, we find that respondents draw very large reputational inferences. Further, we find evidence of both country- and leader-specific reputations: Respondents draw (large) reputational inferences when the past crisis behavior occurred under a different leader — providing evidence of country-specific reputation — and those inferences are even stronger when the past behavior occurred under the current leader, providing evidence of leader-specific reputation. This leader-specific reputation is about half of the size of the country-specific reputation. Finally, we find that leader-specific reputation is more important when the leader is more influential in foreign policy, providing evidence of our unifying theory of influence-specific reputation.

2 A Theory of Influence-Specific Reputation

A reputation is a belief based on an actor's past behavior that informs predictions about their future behavior. Because observers have varying information, and might interpret or weight the same information differently, any actor may hold multiple reputations, though the term usually refers to beliefs about an actor that are common to most observers. Because there are many traits and behaviors about which others can have beliefs, many kinds of reputation are possible (Downs and Jones, 2002). These reputational inferences might concern behavioral tendencies — such as for fulfilling one's threats (Sartori, 2005), for being reliable allies (Crescenzi et al., 2012), or for retaliation (Solnick, 1996) — or traits, such as toughness

(Drezner, 1999, 77) or honesty in communication (Guisinger and Smith, 2002). In some cases, the observers making the inferences are elites (e.g., Mercer, 1996), while in other cases, they are citizens (e.g., Kertzer et al., 2015).

By a large margin, the majority of attention of scholars and policymakers has focused, as we do here, on reputations for *resolve* (Mercer, 1996; Huth, 1997; Tang, 2005; Wolford, 2007; Miller, 2012). A central proposition of theories of deterrence is that building and maintaining a reputation for resolve can deter adversaries (Powell, 1990, Ch. 3) and make compellent threats more credible (Schelling, 1966).[3] And there is a great deal of evidence that policymakers are immensely concerned about their reputations for resolve. In a recent review of reputation in IR, Dafoe et al. (2014, 381) summarized the literature by noting that: "If there is one feature of reputation. . . on which scholars agree, it is that leaders, policy elites, and national populations are often concerned, even obsessed, with their. . . reputation." The roots of that obsession are easy to understand: a reputation for resolve in the eyes of other leaders might deter predation while the same reputation in the eyes of one's citizens might prolong a leader's time in office.

Despite the confident theoretical beliefs of IR scholars and the fervent declarations of statesmen, the record for finding evidence of reputational inferences is mixed. Mercer (1996) used attribution theory to argue that reputations only form when dispositional attributions are made. And despite finding "massive evidence" that leaders think that reputational inferences are being drawn about them, Snyder and Diesing (1977, 187) found "little evidence that statesmen *do* infer an opponent's resolve from his behavior in previous cases." Echoing this, Press (2005) found that while perceptions of interest and power informed assessments of resolve, "past actions" did not (see also Tang, 2005). In the literature on audience costs, a prolonged debate has found mixed evidence of leaders paying at the polls for backing down (see, e.g., Trachtenberg, 2012).

However, there is a growing sense that — as a result of potential methodological biases

[3] Though, in practice, these threats tend to be both rare and rarely successful. See Downes and Sechser (2012).

— these null findings should not be regarded as dispositive. These biases include [1] that common knowledge and unspoken assumptions are less likely to appear in the historical record (Dafoe et al., 2014, 384-385; Weisiger and Yarhi-Milo, 2015); [2] that past actions operate in part through beliefs about *interest* (Weisiger and Yarhi-Milo, 2015); and [3] that strategic selection will, in general, lead observed adverse reputational effects to be biased towards zero (Schultz, 2001). Our studies contribute by *directly* evaluating whether observers draw reputational inferences about the resolve of states, and whether these inferences apply more to leaders or the country.

Influence-Specific Reputations

While scholars have debated the degree to which past actions matter, a related puzzle has gone mostly unscrutinized: do reputations attach most to the leader, the state, or some other entity? This is critical, not just because it is a crucial missing component in theories of reputation, but because it has implications for where we look for evidence of reputations in the first place.

What do we know about the agent-specificity of reputation? Most IR research — 69% by our estimate[4] — has discussed reputations in the context of what we term "country-specific reputations" (CSR; prominent examples include: Walter, 2006; Sartori, 2005; Tomz, 2007b; Crescenzi et al., 2007; Miller, 2012). Country-specific reputations are informed primarily by the country's past actions and adhere to the country as a whole; they are not impacted by leader turnover or other minor changes in decision-making structures, and the past behavior of individual decision makers are not of primary relevance.

A number of recent works have advanced the idea that reputations might be specific to

[4] We coded the bearer of the reputation implied in a sample of articles by selecting the most prominent articles and books in a Google Scholar search of "reputation" within articles citing either Huth (1997) or Jervis (1976). This search strategy returned 42 results, and worked well for returning references that we regarded as central to the IR literature on reputation. 21% of these references had the leader as the bearer of reputation, 69% had the country as the bearer of reputation (62% of which were explicitly country-specific, 38% were implicit), and the remainder unclear or other.

leaders (for examples, see Goemans et al., 2009; Guisinger and Smith, 2002; McGillivray and Stam, 2004; Wolford, 2007; McGillivray and Smith, 2008). Leader-specific reputations (LSR) are reputations informed by the past actions of the leader and adhere just to the leader. Accordingly, leader turnover should lead to a sudden change in reputational beliefs, updated to correspond to background expectations and specific beliefs about the new leader. Folding theories of leader-specific reputations into theories of reputation in IR is necessary to integrate the observation that resolve is likely to vary with individual-level characteristics (Dafoe and Caughey, 2016; Kertzer, 2015), and helps make sense of why new leaders are more resolute in their disputes (Dafoe, 2012, Ch 5) and are more likely to be "tested" by rivals (Wolford, 2007), and why cooperation often resets after leader-turnover (McGillivray and Smith, 2008). Moreover, there is evidence from other contexts that individuals presented with a decision made by a collective use heuristics to punish actors with the most responsibility for a given decision, implying the existence of individual-level reputations in laboratory experiments (Duch et al., 2015). Publics are also able to make these judgments, punishing leaders responsible for bad war outcomes but sparing those who were not in charge initially (Croco, 2011).

What are the observable implications of these different kinds of reputation? Leader-specific reputations will suddenly and dramatically change with leader-turnover, whereas country-specific reputation will be unaffected by leader-turnover. Putting these together, we see that we can identify country-specific reputation by looking in settings where there is leader turnover, and we can identify both country- and leader-specific reputation by looking at settings where there is no leader turnover. Our experimental design does this.

Our theory and analysis builds from a simple dichotomy between the leader and the country. Of course, reputations could, in principle, adhere to any kind of agent—such as a ruling group, a political party, an organization, a clan, or an army—or institutional characteristic—such as a component of a constitution, an organization culture, an electoral system, or even a religion. We regard our operationalization as a productive first approximation that reflects the main approaches in the IR literature. However, it is worth remembering that "the country" here is actually a bundle of agents, institutions, and temporally-correlated factors, each of which could be the target of a reputational inference. Our approach collapses the reputations of all entities that tend not to change with leader-turnover into *country-specific reputation*. Future research should further theorize and empirically distinguish the many possible bearers of reputation.

The logic of our theory is that inferences are most useful when they are most accurate, and most accurate when they correspond most closely to the decision making process in question. Leaders and citizens who wish to draw the most useful inferences about others will strive to understand the decision-making influence of actors and use that information to formulate their inferences about reputations. In regimes where a single leader holds most of the decision-making power, it would be most sensible for observers to draw leader-specific reputational inferences; in regimes where the leader's influence is heavily constrained by elites, electoral incentives, a political party, a parliament or other country-specific institutions, country-specific reputational inferences would be relatively more useful. This analytical distinction can also be made conditional on the issue-area; for example, reputational inferences about democracies should be more leader specific in domains where democratic leaders are less constrained, such as national security, compared to other domains, such as monetary policy.[5]

We formalize our main theoretical predictions using a canonical mathematical model of reputation (summarized here; Appendix §B provides more details). We begin with a two-period game of sequential move chicken, also known as the chain store game (Selten, 1978; Kreps and Wilson, 1982; Milgrom and Roberts, 1982). The defender can have the usual chicken payoffs, in which case the defender prefers to back down rather than resist a challenge, but most prefers for the other side not to challenge in the first place. Alternatively,

[5] Our theory has a parallel in the literature on retrospective voting, where attributions of responsibility for economic performance are theorized to "strongly reflect the nature of policymaking in the society and the coherence and control the government can exert over that policy." See Powell and Whitten (1993, 398).

the defender can be *intrinsically honorable*, in which case the defender prefers to resist any challenge rather than back down, but again they most prefer for the other side not to challenge in the first place. The literature often refers to our "intrinsically honorable" type as "tough," "crazy," or "extreme." We use the term "intrinsically honorable" because we believe the concept of honor better represents the logic of resistance to coercion.

The central result from this literature (Mailath and Samuelson, 2006) is that when agents engage each other multiple times, they are sufficiently patient, and there is some sufficient probability of intrinsically honorable types[6], then some agents will act as if, and thereby develop a reputation for being, intrinsically honorable. Rational observers will draw reputational inferences, and some potential challengers will avoid challenging honorable agents (agents who behave as if they are intrinsically honorable), though they will challenge agents who have revealed themselves to not be honorable.

We denote the reputational inference as θ. Formally, θ is the change in the conditional probability that an agent will resist a challenge (will be resolved), depending on whether the agent backed down in the previous round or stood firm:

$$\theta = P(\text{Agent Stand Firm Now}|\text{Agent Stood Firm Before})$$

$$-P(\text{Agent Stand Firm Now}|\text{Agent Backed Down Before})$$

For rational observers, all else equal, θ will also be the effect of past resolute behavior on their perception of the agent's resolve. In the experimental set-up that follows, θ is identified as the difference between the two conditions relating to an agent's past actions; the difference between when they stood firm in the past and when they backed down. To the extent that $\theta > 0$, then we have found evidence that reputations for resolve form based on past resolved behavior.

Following the observable implications we outlined above, the country-specific reputation

[6] Or there is sufficient uncertainty about the length of the game or whether the game can go on forever.

can be identified by looking at the effect of past actions when there is leader turnover. To the extent that the reputation attaches to the country, observers should draw inferences from past behavior even when there is leadership turnover. We denote the country-specific effect of past actions, for a given level of leader influence X, as $\theta_{CSR,X}$, which is equal to the effect of past actions when there is a different leader (DL): $\theta_{CSR,X} = \theta_{DL,X}$

> H_{CSR} : *even when there is leader turnover, observers will draw reputational inferences across periods.* $\theta_{DL,X} > 0 \quad \forall X \in \{LI, HI\}$.

If reputations adhere somewhat to leaders, then the effect of a state's past actions (θ) should be weaker when there is leader turnover, compared to no leader turnover. Formally, for a given level X of leader influence, we would find support for leader-specific reputation if $\theta_{LSR,X} = \theta_{SL,X} - \theta_{DL,X} > 0$, where SL denotes that the **S**ame **L**eader is in power across periods. If reputation only adheres to leaders, then $\theta_{DL,X} = 0$, so $\theta_{SL,X} - \theta_{DL,X} = \theta_{SL,X}$.

> H_{LSR} : *at least when leader influence is high, observers will draw stronger reputational inferences across periods when the leader is the same, compared to when the leader is different.* $\theta_{LSR,HI} = \theta_{SL,HI} - \theta_{DL,HI} > 0$

Finally, influence-specific reputation can be defined as the proposition that leader-specific reputation will be greater for leaders with high influence (denoted HI), compared to those with low influence (denoted LI): $\theta_{LSR,HI} > \theta_{LSR,LI}$. We summarize these hypotheses about country-specific reputation (H_{CSR}), leader-specific reputation (H_{LSR}), and influence-specific reputation (H_{ISR}) here. These hypotheses follow from the formal model in Appendix B.

> H_{ISR} : *leader-specific reputation will be greater for high values of leader influence.* $\theta_{LSR,HI} > \theta_{LSR,LI}$

3 Research Design

The overwhelming majority of recent work on reputation uses observational data, either in case studies that examine decision-makers' deliberations in crises or in large-N studies that look for the effects of reputations using COW/MID data. Each approach requires significant trade-offs, and we see the experimental approach as one part of a larger attempt to triangulate research on reputations using a combination of methods that build upon one another.[7] We also note that — contrary a prevailing perception — even scenario-based survey experiments can suffer from problems of internval validity similar to the confounding problems that plague observational studies (Dafoe et al., 2015). We explain our approach to addressing these issues in Appendix §C.

The experiments described below were conducted on samples of subjects drawn from Amazon's Mechanical Turk (MTurk) labor marketplace, an increasingly popular resource for experimental social science, particularly when the study design does not require a physical presence in the laboratory. Berinsky et al. (2012, 366) show that mturk samples are "often more representative of the general population and substantially less expensive to recruit" than the "convenience samples" typically used in political science.[8] They also demonstrate the ability to replicate results from nationally-representative samples — e.g., Tversky and Kahneman's (1981) classic "asian disease problem" — using mturk workers.[9] In keeping with "best practices," we limited participation in the study to mturk workers located in the United States, who had completed $\geq$ 50 HITs, and whose HIT approval rate was >95%, and re-fielded the survey each morning of the days it was active, giving us several successive "waves."[10]

[7] For similar approaches on status and resolve, see Renshon (2017) and Kertzer (2015).

[8] Though compared to nationally representative samples, mturk workers tend to be younger and more ideologically liberal.

[9] For more on this, see Rand (2012). For a different viewpoint, see Krupnikov and Levine (2014), though their caution applies particularly to mturk studies that require subjects to read a significant amount or trust information from an experimenter; two attributes that were not important in the studies described here.

[10] Different waves did not themselves correlate — either by themselves or in interaction with the treatments — with any of the outcomes of interest in either study.

4 Study 1: Reputation in a Hypothetical Scenario

Study 1 was fielded over the course of four days in August 2014. Subjects (N=1804) were recruited via MTurk and the study was administered on the Qualtrics survey platform. This sample was highly educated (49% had at least a 4-year college degree), well-balanced on gender (41% female), contained far more Republicans (30%) than one would find in a student sample, and the mean age was 32.[11]

Subjects were allowed to participate only once and were paid a base rate of $0.60. We incentivized attention and effort through bonus payments of $0.40, which subjects could earn by answering several "memory check" questions at the end of the study correctly.[12] We included these incentives because the empirical literature is very clear that small to moderate incentives can make a large difference in attention and effort (Hertwig and Ortmann, 2003).[13]

Term:	Definition:	Hypothesis:
Effect of past actions (past resolved behavior)	θ	$\theta > 0$, perceived resolve will be greater when **A** stood firm in the past than when they "backed down"
Country-Specific Reputation (CSR)	$\theta_{CSR,X} = \theta_{DL,X}$	$\theta_{DL,X} > 0$, $X \in \{LI, HI\}$, the effect of past actions (θ) will exist when there is leader turnover.
Leader-Specific Reputation (LSR)	$\theta_{LSR,X} = \theta_{SL,X} - \theta_{DL,X}$	$\theta_{SL,HI} - \theta_{DL,HI} > 0$, when leader influence is high, the effect of past actions will be greater when there is no leader turnover (compared to θ when there is leader turnover)
Influence-Specific Reputation (ISR)	$\theta_{LSR,HI} - \theta_{LSR,LI}$	$\theta_{LSR,HI} - \theta_{LSR,LI} > 0$ $\implies (\theta_{SL,HI} - \theta_{DL,HI}) - (\theta_{SL,LI} - \theta_{DL,LI}) > 0$, leader-specific reputation will be greater when leader's influence over policy-domain is high (compared to low)

LI = Low influence, HI = High influence
DL = Different leader, SL = Same leader

Table 1: Recap of terminology and hypotheses

In the introduction to the study, subjects were notified about the memory check, and then told that they would be asked to read text describing a scenario about two countries engaged in a territorial dispute (labeled Country **A** and Country **B** for purposes of generality). The

[11] Demographic characteristics are summarized in Appendix §D.

[12] Bonuses implemented with `MTurkR` package.

[13] However, the relationship between incentives and performance does not appear to be entirely monotonic when incentives are extremely high. See Ariely et al. (2009).

survey vignette is contained in Appendix §E. In the survey, subjects were randomly assigned to conditions relating to:

1. *Past Actions*: whether or not Country A stood firm or backed down in previous crises
2. *Leadership*: whether the previous crises occurred under either the current or previous leader of Country A
3. *Influence*: whether the leader of Country A exercised "complete" or "very little" control over foreign policy
4. *Power*: whether Country A had significantly more, less, or equal military power relative to Country B

This resulted in a 2 (past actions) × 2 (leadership) × 2 (influence) × 3 (power), fully crossed over experimental design, illustrated in Figure 1.

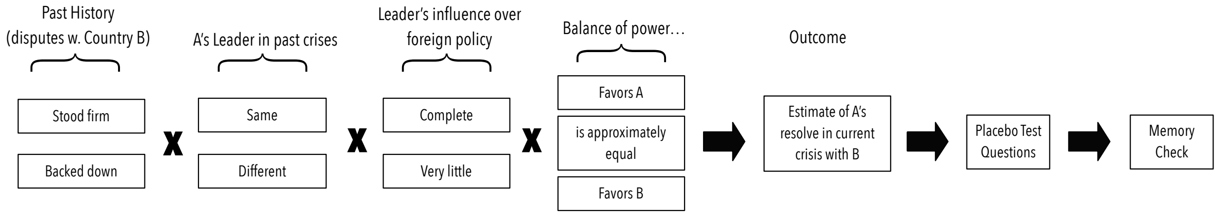


Figure 1: **Design of Study 1**

After the initial survey vignette, subjects were given a reminder/summary of the salient aspects of the crisis, and asked to rate the likelihood that Country **A** would back down: "What is your best estimate, given the information available, about whether Country A will back down in this dispute?" Answers were scaled from 1 ("Country A is **very likely** to back down [80% to 100% chance]") to 5 ("Country A is **very unlikely** to back down [0% to 20% chance]"), where a "5" represented the greatest estimate of A's resolve in the current crisis.

After our outcome question, we asked a placebo question focused on the democratic-ness of the countries in the scenario. In particular, we asked how democratic subjects estimated Country **A** to be, on a scale from -10 (fully autocratic) to 10 (fully democratic). Finally, subjects were asked four manipulation/memory check questions (1 for each of the randomized treatments) to assess their attention level, and a battery of demographic questions and dispositional/ideology scales.[14] Our manipulation and memory check questions confirm that the respondents noticed and remembered our treatments: the overwhelming majority (82%) were able to reproduce all four experimental treatments correctly, and 95% were able to remember at least three out of four treatments correctly.[15]

Does Past Resolved Behavior Affect Estimates of Resolve?

Do reputations based on past behavior exist at all, or are estimates of resolve based entirely on the current balance of capabilities and interests (à la Press 2005)? Our results, both with and without individual-level controls, show a large, statistically significant, effect of *Stood Firm*: standing firm in the past increases estimates of **A**'s resolve by 45 percentage points, more than doubling, from 35% to 80%.[16] This quantity is θ, and represents the difference in subjects' estimates of **A**'s resolve in the "stood firm" and "backed down" conditions.

Figure 2 displays the distributions and means (vertical lines) by *Stood Firm* treatment condition.[17] Subjects exposed to our *Stood Firm* treatment gave far higher estimates of **A**'s resolve in the current crisis ($t(1802) = -48.32$, $p < 0.001$). This sizable difference is, if anything, biased downwards by ceiling effects, as the distribution under *Stood Firm* is

[14] Our placebo test is reproduced in Appendix §F, manipulation checks are listed in Appendix §G and all demographic scales and dispositional measures are listed in Appendix §H.

[15] Only 13 out of 1,804 (0.7%) subjects answered none of the memory check questions correctly.

[16] There are two ways to look at our responses: either as a scale (between 1-5) or as an estimated probability that Country A will back down. We discuss our results in the latter form to aid in interpretability. Since each number on the scale corresponded to a probability range (e.g., 2 corresponds to the range, $20-40\%$), when we plotted the results, we simplified by taking the midpoint of each range, so that "1" was equal to 10%, 2 was equal to 30%, etc. As a result of this procedure, the maximum estimated resolve (a "5" on a scale from 1-5) is described on the plots as 90%.

[17] Results are substantively identical if we control for individual-level covariates or other experimental conditions. See Figure 13 in Appendix §K.

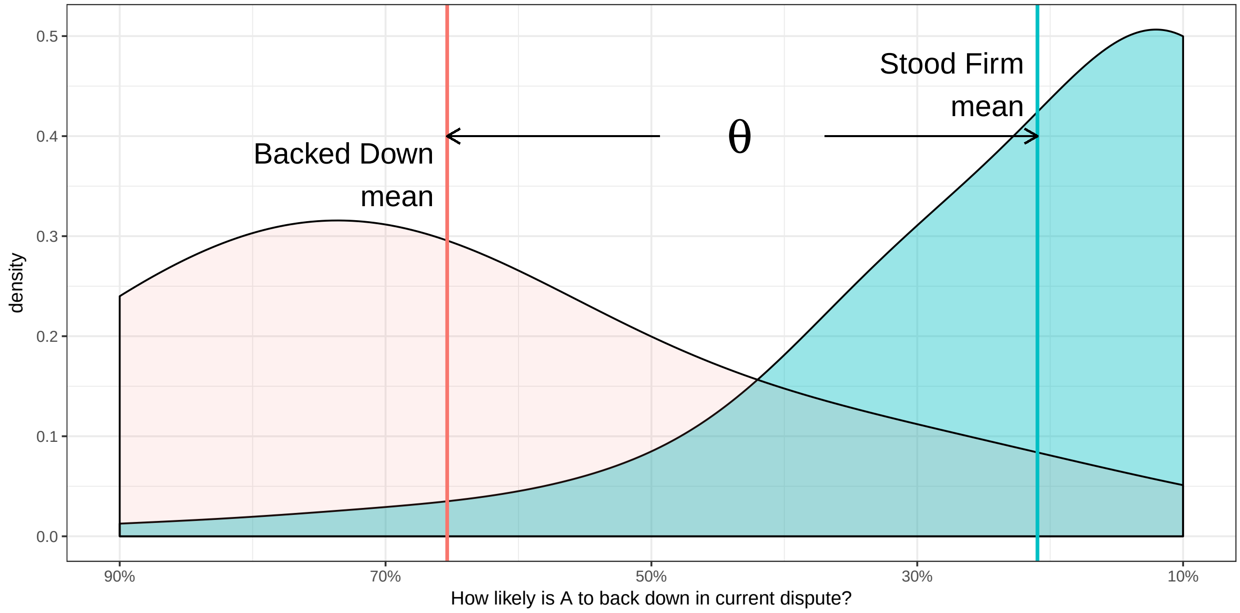


Figure 2: θ**, The effect of past actions on reputations for resolve**: Treatment conditions: **A** either backed down in the last 2 disputes (N=906) or stood firm in last two disputes (N=898).

heavily weighted on the highest category of resolve.

The balance of power between **A** and **B** in the scenario did not affect the development of **A**'s reputation. Whether **A** was more, less or equally as powerful as **B** in the present, **A**'s past behavior significantly impacted subjects' estimates of their resolve in the present.[18]

Country and Leader-Specific Reputations

Next we begin to "unpack" reputations by asking how much of subjects' reputational inferences adhered to (a) the country itself or (b) the country's leader. We first estimate θ_{CSR} (from Equation 2 in Appendix B), which is the effect of past actions when there is a leadership change. Given the leadership change, any effect of past actions on perceptions of current resolve cannot be attributed to leader-specific reputation. We find large positive effects of past actions: $\hat{\theta}_{CSR} = 35\%(\pm 3\%)$.[19]

[18] Both the magnitude and statistical significance of our estimate of θ remain constant across power conditions. See Table 4 in Appendix J.

[19] To estimate this, we included an interaction, *Stood Firm* × *Same Leader* along with the main treatment conditions. For regression results, see Table 3 (in Appendix §I). For a plot that isolates only Country-Specific

Having found support that reputations adhere to countries (or to be precise, do not completely disappear with leader turnover), we turn to evaluating whether some reputation also adheres to the leader. Leader-specific reputation can be estimated by looking at how the effect of past actions increases when the leader is the same, as compared with when the leader is different (and for an easy test of LSR we set leader influence to high): $\theta_{LSR,HI} = \theta_{SL,HI} - \theta_{DL,HI}$.[20]

Figure 3 presents our results of leader-specific reputation. This plot depicts $\theta_{SL,HI}$ (the effect of past actions) by *Same Leader* condition (when *High Influence* $= 1$), along with 95% confidence intervals. It is generated by estimating the regression in Column 2 (a-b) of Table 3, and then using CLARIFY to generate "first differences," by holding covariates at their mean or median and switching *Past Actions* from "backed down" to "stood firm" in both leader treatments. The y-axis represents the *increase* in subjects' estimates of **A**'s resolve. The first, left, estimate is the effect of past actions under a different leader, and is the country-specific reputation (CSR) we just analyzed.[21] The effect of past actions when *Same Leader* $= 1$ is on the right, and is estimated to be 51%, a significant increase over the CSR at 28%. The difference between these two quantities is estimated to be a full 23 percentage points on the y-axis. We thus find strong support for the notion that reputations can adhere to individual leaders as well as to states. In fact, if we make the test harder, by including only conditions in which the leader is described as having "very little" influence over policy, we still find statistically significant evidence in favor of leader-specific reputations. In effect, reputations adhere to leaders even when there is common knowledge that they had very little influence; far from struggling to gain reputations, leaders cannot escape them.

Reputation, see Figure 14 in Appendix §L.

[20] Equation 3 in Appendix B.

[21] Technically, this (28%) is the estimate of the CSR when *High Influence*=1, which is estimated to be lower than the same quantity when *High Influence*=0 (when it is 35%).

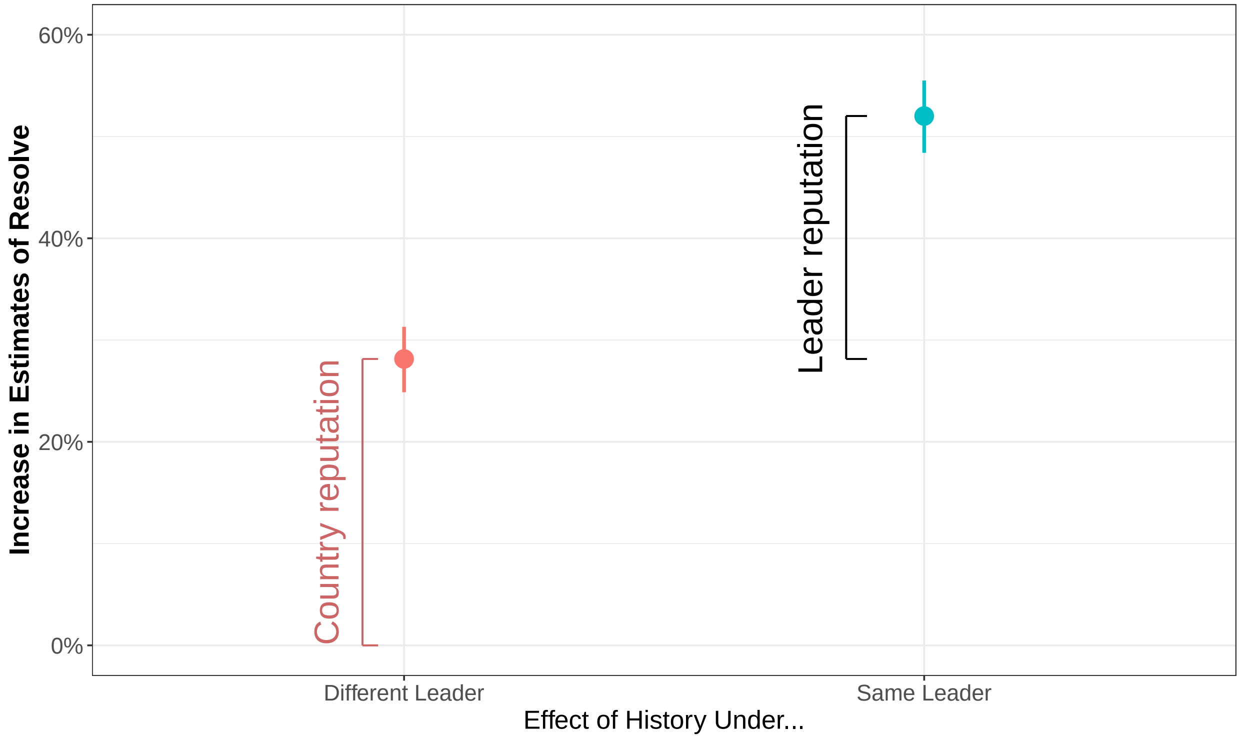


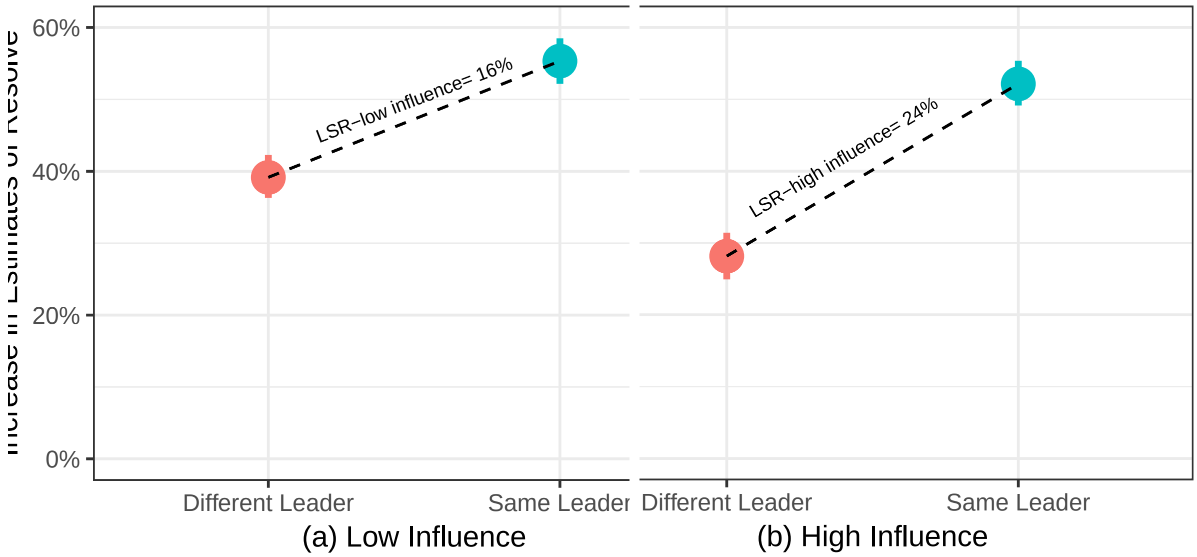
Figure 3: **Leader-Specific Reputations**

Influence-Specific Reputations

Thus far we've shown that reputations adhere to both states and to leaders, providing estimates of the extent to which each contributes to observers' overall reputational inferences. We turn now to evaluating our theory of *Influence-Specific Reputation* which provides a parsimonious unifying framework for understanding when leader-specific and country-specific reputations will be relatively more or less important. To do so, we extend the previous analysis by considering how the magnitude of leader-specific reputations changes depending upon the influence the leader is perceived to have. Recall that our theory of *Influence-Specific Reputation* predicts that leader-specific reputation will be greater when the leader has high, rather than low, influence: $\theta_{LSR,HI} > \theta_{LSR,LI}$. Accordingly, Figure 4 compares the effect of LSR for high and low influence.[22] Figure 4 shows evidence consistent with our

[22] Results were calculated from the regression reported in Table 3, Column 3 (a-b). This set of models include a three-way interaction, *Same Leader* × *Stood Firm* × *High Influence*, as well as all lower-order interactions. These regressions (and interaction terms) are non-parametric since our variables are all indicator

theory of Influence Specific Reputation: leader-specific reputation increases from 16% under low influence to 24% under high influence, increasing the size of leader-specific reputation by a striking 50% (or 8 percentage points (±6%, $p < 0.05$).[23]



(a) Low Influence

(b) High Influence

Figure 4: **Influence Specific Reputations**: Leader-specific effect by *Influence* condition. (a) "low influence" and (b) "high influence."

Our abstract scenario-based survey experiment provided several important results. Countries acquire reputations that persist across leaders, but those same leaders also acquire reputations which are more important when they are influential. All of these reputational inferences are substantively large. However, this study is based on a highly abstract scenario (Country **A** vs **B**) with little context and information. It may be that reputational inferences form in such an informationally impoverished setting, but not in richer, more realistic settings. To address this possibility, we turn now to our second study.

variables for a specific experimental manipulation.

[23] Estimating this with confidence is difficult because it is a second-order interaction. Nevertheless, we do find statistically significant results, $p_{one-sided} < 0.01$.

5 Study 2: Reputation in a U.S.-Iran Confrontation

Study 2 presented respondents with a realistic conflict between the United States and Iran.[24] Fielding this realistic follow-up study offers several advantages. First, replicating any study on a new and different sample (subjects from Study 1 were excluded from Study 2) should increase our confidence in the results. Doing so in a study that is conceptually similar but differs on important details does so further since it is correspondingly less likely that the results are being driven by an idiosyncratic feature of our design. In addition, the fact that our same predictions are found in a second study, employing the same conceptual design and analysis strategy, reduces the risk (to the reader) that our predictions were formulated after seeing the results.

Second, a rich realistic scenario provides a harder test for theories of reputation. In the abstract scenario, little information was provided, so it is less surprising if respondents cue off of the information from our treatments. By contrast, the US-Iran scenario is detailed and realistic, and should hold fixed respondents' beliefs about other features of the scenario, since the details will either be specified or the respondents can call upon their pre-existing beliefs about these countries to inform their understanding of the scenario. The results from our placebo test, discussed later, are consistent with this. This greater detail and realism means that respondents will have more firm beliefs about power, interest, and other factors that influence observers' judgement about resolve in the real world, but that are absent in an abstract scenario. In particular, our realistic scenario provides more opportunity for partisan or other ideological beliefs about a crisis involving Iran (which might be dominant in the real world) to express themselves and overpower the informative effects of past actions. In summary, in addition to offering a replication on a new sample and design, Study 2 should provide a harder test of our hypotheses and yield more realistic, externally valid, results.

The scenario itself was inspired by a real war-game led by the Brookings Institution

[24] It was fielded over six days in early February 2015 (before the comprehensive blueprint for a nuclear agreement was announced) and was identical in general structure to Study 1.

(Pollack, 2012) which involved a scenario in which: [1] the U.S. and Iran are engaged in negotiations focused on the latter's nuclear program; [2] Iran has withdrawn from nuclear talks and blamed the U.S. for covert activities, including a bombing at a nuclear facility; [3] Iran-backed terrorists retaliated by setting off a bomb that killed U.S. tourists and nuclear scientists in another country; [4] the U.S. has blockaded the Straits of Hormuz.

As before, subjects were drawn from an MTURK sample, and randomly assigned to conditions relating to *Past Actions*, *Leadership*, *Influence* and *Power*.[25] One minor change from Study 1 is that we revised the power conditions to more plausible levels for the US-Iran dyad; we now assign subjects in Study 2 to conditions in which Iran either has "significantly less" or "slightly inferior" capabilities relative to the United States. The necessity of changing the study, even in this minor way, highlights the difficulty with using more realistic scenarios: subjects are likely to have strong priors concerning the state of the world, making some manipulations implausible. All treatments were manipulated independently, and the order in which they were displayed to subjects was randomized.[26]

Influence-Specific Reputations in a Richly Detailed Scenario

In Study 1, standing firm in the past increased estimates of **A**'s resolve by 45 percentage points (from 35% to 80%). Here, in the context of a real conflict between the U.S. and Iran, our *Past Actions* treatment increased respondents' beliefs that Iran would stand firm by 34 percentage points (from 32% to 66%; see top panel of Figure 5). As before, this was statistically significant with or without demographic controls.[27]

In Study 1, the country-specific reputation — the effect of standing firm (θ) when the leader was different — was positive and statistically significant, indicating that reputations can and do accrue to countries (as distinct from leaders). There, we estimated past resolute

[25] Study 2 is reproduced in Appendix §M.

[26] The exception to this are the *Past Actions* and *Same Leader* treatments, which were always displayed next to one another, for ease of reading.

[27] Full results from regression models are contained in Table 3 in Appendix §I.

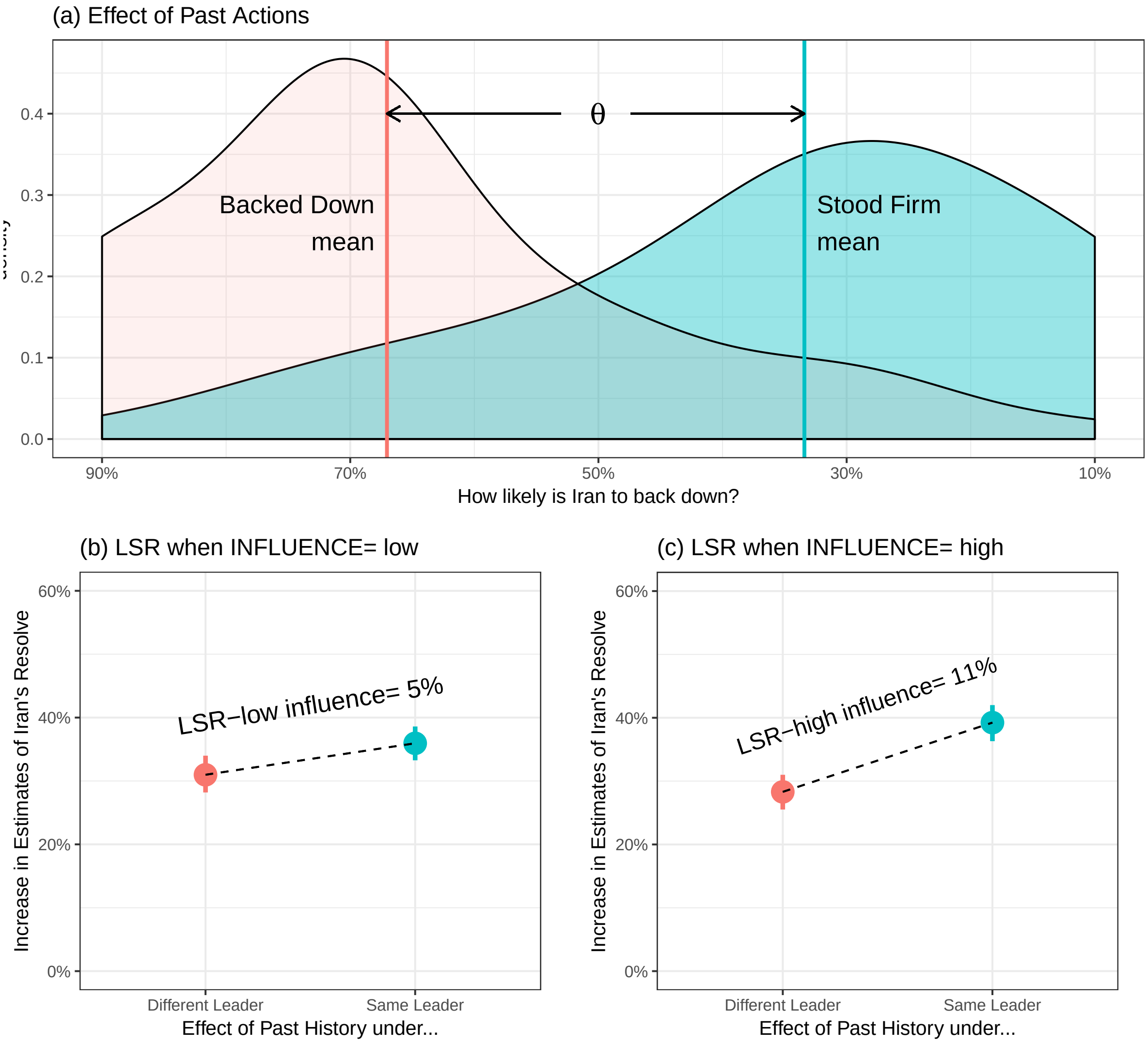


Figure 5: **Iran Scenario Results**: Results show (a) the effect of *Past Actions* on estimates of Iran's resolve; (b) leader and country-specific reputations for Iran. Panels (c) and (d) show the magnitude of the leader-specific reputation under conditions of *Low* and *High* influence.

behavior to increase estimates of resolve by 35% when the leader was different. In Study 2, our estimate of the CSR was nearly identical: the effect of standing firm when there was leadership turnover was an increase in perceptions of resolve of 38%.

In Study 1, we also found evidence for leader-specific reputation, which is defined as the difference between θ when the leader is the same versus when it is different. In Study 2 we again found strong evidence for LSR, depicted in the bottom right panel of Figure 5 where *Leader Influence* is set to High. The effect of past actions is estimated to be 10% greater when there is no leader-turnover, providing evidence of leader-specific reputation.[28]

Lastly, we evaluate our unifying theory of *Influence Specific Reputations*. Our theory of influence-specific reputation states that LSR will be larger when the leader has greater influence over the policy in question than when they do not: $\theta_{LSR,HI} > \theta_{LSR,LI}$. Again, we find strong support for our theory of influence-specific reputation in Study 2 which reveals that changing leader influence more than doubles the effect of leader-specific reputation (an increase of 6 percentage points, from 5% to 11%, $p < .05$).

Discussion

Realism of Scenarios

We consider here several possible issues in the interpretation of our results: could our simplified scenarios pose a threat to external validity (for a more detailed discussion of this issue, see Appendix §A.1)? Might our experiments present the causal factor of interest (the treatment) in an implausibly salient manner, thus leading us to over-estimate the effect? Or could it be that our experiment presented the treatment in an overly abstract and digested format, or in a manner that permitted the respondent to guess about our purpose, thus potentially cueing the respondent to draw the reputational inference that we sought?

Such concerns about the mapping between experimental effects and real-world effects are

[28] The effect is substantively smaller than the 24% LSR effect that we observed in Study 1.

important, but despite these concerns we believe that our results remain informative. First, given the difficulty of causal inference in IR, scholars should be primarily concerned with estimating the existence and sign of effects, rather than their magnitude. Concerns about over-estimating the size of our effect are thus far less pressing than the question of whether we have identified an effect in the first place, given the state of our collective knowledge. Second, there are other reasons (low stakes, an inattentive and non-expert audience) for believing that our estimates are actually *smaller* than their real world counterparts. Third, our US-Iran scenario was included precisely to add crucial realism; strikingly we found similar results even though the treatment text was a small portion of the vignette (the influence treatment, for example, involved a mere 5% of the words of the total vignette). Finally, it is implausible that respondents would have inferred the ISR theory that we were studying in both scenarios given that ISR involves second-order interactions and the US-Iran scenario was richly detailed.

Future research should try to pin down plausible estimates of the magnitude of real-world effects, perhaps by designing more realistic and subtle ways of communicating the influence of the foreign decision-maker. Similarly, we should keep in mind a crucial scope condition: our theory only predicts that reputations should be influence-specific in domains where the observers can perceive the level of influence of the decision-maker with sufficient precision. Nevertheless, we are able to learn much from reliably estimating the sign and existence of an effect, as we have done here.

Samples and Inferences

We consider here questions concerning how our sample influences the inferences that we can draw. Specifically, (1) does our sample of the public allow us to make plausible inferences about elites? (2) Is our ISR theory, which is based on rationalist observers, plausible for members of the general public? For more detailed discussion, see Appendix §A.2.

Do our findings generalize to elites? While an ideal study would have also surveyed the perception of elites and leaders, such samples are extremely costly; large-n samples of actual leaders are near-impossible to acquire. For this reason, research programs should not *begin* on elite samples, but rather progress to them over time and after replications have corroborated preliminary findings (McDermott, 2002).

The primary issue here is the extent to which the causal processes that generate the effect are likely to vary across populations, and if so, in what ways. Several recent studies that have examined differences between elites and the general population have found little to distinguish them (for references, see §A.2), providing evidence that at least some causal processes do not vary much across these populations. The key difference we envision is that elite observers should be more rational in their assessments compared to the public; they are more informed, have been selected for and highly trained in rational reasoning, and are closer to cultures that encourage rational assessment of foreign policy threats relative to members of the public. These differences suggest that the public should be less likely to exhibit the subtle rationality of our ISR predictions and that any effects we see are likely attenuated relative to the effects that would arise with an elite sample.

Is it reasonable to expect the public to draw rational inferences? There is certainly abundant evidence of members of the public exhibiting deviations from rationality and having only limited information about foreign affairs. However, in many domains humans also appear to learn in an approximately rational way (Holyoak and Cheng, 2010) given their bounded information. In particular, inferences about the resolve of individuals and groups is something that is valuable in our daily lives and was valuable in our ancestral environment. Thus, it is plausible that society and evolution have endowed us with faculties sufficiently effective at drawing reasonable inferences about reputation in inter-group conflict. Future work should examine how imperfect information, or particular heuristics and biases, could lead to deviations from ISR.

Conclusion

Despite the prominence of reputations in IR research, one prominent question has yet to be addressed: whether reputations adhere to leaders, states or other entities. We innovated by offering a unifying theory of *influence-specific reputation* that explains when reputations are most likely to adhere to a specific entity. Our theory also makes systematic predictions about how the agent-specificity of reputation will vary by regime type, policy domain, and any other feature of the country or dispute. This theory was based on the simple premise that observers are approximately rational, and thus draw reputational inferences that are most useful to them, using their knowledge about decision making influence. These predictions, as well as the core concepts of CSR, LSR, and ISR, were shown to emerge from a mathematical model of reputation.

Because of the severe inferential difficulties associated with observational data, we designed two scenario-based survey experiments to tease apart the agent-specificity of reputations. In the first scenario, we manipulated the past actions of **A** (whether they "stood firm" in past disputes), whether a leadership turnover had occurred, and the level of the leader's influence over policy, along with the power differential in the current dispute. We then asked subjects to estimate **A**'s resolve in the current crisis. The second study was similar in structure, but centered on a detail-rich and realistic scenario of a conflict between the U.S. and Iran. Across both studies, we found evidence that reputations develop based on past resolved behavior, and that reputations adhere to both states and leaders. We also found support for our theory of influence-specific reputations: leader-specific reputations were stronger when the leader was more influential in foreign policy.

One issue that could be explored in the future is the correlation in preferences across leaders; in systems where leaders are more like each other in their preferences, rational reputational inferences should increasingly adhere to that common unobserved factor of leader preferences, and thus look like CSR. At the limit, if leader preferences are close to perfectly correlated then LSR should become small or even vanish.[29] Future work could also investigate the extent to which these results generalize to other populations, such as elites or the public in other countries. We expect that elites would exhibit even stronger ISR effects, given that their beliefs and reasoning processes are likely closer to rational; however, there are other possibilities, such as if elite knowledge crowds out reputational inferences or if elites believe that crises are less correlated in their preference structure across time than the public does. Future work could also investigate publics from other countries, which may have different beliefs about reputation or decision-making influence, as would arise if publics erroneously attributed to other countries similar domestic political processes as their own.

Our findings regarding the presence and complexity of reputational inferences offer important contributions to at least three larger scholarly literatures. First, our results shed light on the debate within deterrence theory on the role of reputations in assessments of credibility. Our experimental findings on the presence and substantively large size of both country- and leader-specific reputations provide new and compelling evidence in support of the position that reputational inferences are likely to be central factors shaping the credibility of deterrent threats. And by providing micro-level evidence of reputational inferences being drawn by individuals in a variety of situations, our findings complement recent large-N statistical studies (e.g., Weisiger and Yarhi-Milo, 2015) that find empirical patterns consistent with reputations forming on the basis of past behavior. Our theory also provides an explanation for some patterns in IR. Three studies providing evidence of leader-specific reputation (McGillivray and Stam, 2004; McGillivray and Smith, 2004, Dafoe, 2012, Ch 5) found that the results were stronger for countries with more authoritarian governments. These findings are consistent with our theory of ISR.

Finally, our theory and findings regarding leader- and influence-specific reputations point toward the utility of a focus on individual leaders as central units of analysis. More broadly, they are suggestive of a need for scholars to think about and pursue more contingent theories

[29] It need not vanish, however, if reputation is based on strategic uncertainty. For example, reputational inferences arise in the infinitely repeated Prisoner's Dilemma, even when there is complete information.

of deterrence and international conflict in which different levels of analysis are called for depending upon how much control high-level decision makers exert over security decisions. While we focus on reputations for specific leaders, we can extend the underlying logic to argue that when decision-makers exert strong control over policy choices, we must place that individual at the center of theory-building efforts. We would expect other actors to not only draw reputational inferences about that leader but to also be sensitive to their other traits or characteristics. This does not inherently require a psychological/cognitive approach, but it does suggest that this central decision maker becomes the theoretical focal point around which the larger strategic environment should be situated.

References

Alatas, V., L. Cameron, A. Chaudhuri, N. Erkal, and L. Gangadharan (2009). Subject pool effects in a corruption experiment. *Experimental Economics 12*(1), 113–132.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar (2009). Large stakes and big mistakes. *The Review of Economic Studies 76*(2), 451–469.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating online labor markets for experimental research. *Political Analysis 20*(3), 351–368.

Crescenzi, M. J., J. Kathman, K. Kleinberg, and R. Wood (2012). Reliability, Reputation, and Alliance Formation. *International Studies Quarterly 56*(2), 259–274.

Crescenzi, M. J., J. Kathman, and S. B. Long (2007). Reputation, history, and war. *Journal of Peace Research 44*(6), 651–667.

Croco, S. E. (2011). Leader culpability, war outcomes, and domestic punishment. *American Political Science Review 105*(3), 457–477.

Dafoe, A. (2012). *Resolve, Reputation, and War: Cultures of Honor and Leaders' Time-in-Office.* Ph. D. thesis, University of California, Berkeley.

Dafoe, A. and D. Caughey (2016). Honor and war: Southern U.S. presidents and the effects of concern for reputation. *World Politics 68*(2).

Dafoe, A., J. Renshon, and P. Huth (2014). Reputation and status as motives for war. *Annual Review of Political Science 17*, 371–393.

Dafoe, A., B. Zhang, and D. Caughey (2015). Confounding in survey experiments. Manuscript. http://www.allandafoe.com/confounding.

Downes, A. B. and T. S. Sechser (2012). The illusion of democratic credibility. *International Organization 66*(3), 467–489.

Downs, G. and M. Jones (2002). Reputation, compliance, and international law. *The Journal of Legal Studies 31*(1), 95–114.

Drezner, D. W. (1999). *The sanctions paradox.* New York, NY: Cambridge University Press.

Duch, R., W. Przepiorka, and R. Stevenson (2015). Responsibility attribution for collective decision makers. *American Journal of Political Science 59*(2), 372–389.

Fearon, J. D. (1994). Domestic political audiences and the escalation of international disputes. *The American Political Science Review 88*(3), 577–592.

Fudenberg, D. and J. Tirole (1995). *Game Theory*. Cambridge, MA: MIT Press.

Goemans, H. E., K. S. Gleditsch, and G. Chiozza (2009). Introducing archigos: A data set of political leaders. *Journal of Peace Research 46*(2), 269–283.

Guisinger, A. and A. Smith (2002). Honest threats. *Journal of Conflict Resolution 46*(2), 175–200.

Hafner-Burton, E. M., B. L. LeVeck, D. G. Victor, and J. H. Howler (2014). Decision maker preferences for international legal cooperation. *International Organization 68*(4), 845–876.

Hertwig, R. and A. Ortmann (2003). Economists' and psychologists' experimental practices. In I. Brocas and J. Carrillo (Eds.), *The Psychology of Economic Decisions*, pp. 253–272. Oxford University Press.

Holyoak, K. J. and P. W. Cheng (2010). Causal learning and inference as a rational process. *Annual review of psychology 62*, 135–163.

Huth, P. K. (1997). Reputations and deterrence: A theoretical and empirical assessment. *Security Studies 7*(1), 72–99.

Jervis, R. (1976). *Perception and Misperception in International Politics*. Princeton, NJ: Princeton University Press.

Jervis, R. (1982). Deterrence and perception. *International Security 7*(3), 3–30.

Kertzer, J. (2015). Resolve in international politics. Book manuscript.

Kertzer, J. D., J. Renshon, and K. Yarhi-Milo (2015). How do observers assess resolve? Working Paper. http://jonathanrenshon.com/Site/CurrentResearch_files/KertzerRenshonYarhi-Milo021415.pdf.

Kreps, D. M. and R. Wilson (1982). Reputation and imperfect information. *Journal of Economic Theory 27*, 253–279.

Krupnikov, Y. and A. S. Levine (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science 1*(1), 59–80.

Linde, J. and B. Vis (2016). Do politicians take risks like the rest of us? *Political Psychology*.

Loewen, P. J., L. Sheffer, S. Soroka, S. Walgrave, and T. Shaefer (2015). Are Politicians Better Decision Makers? Working Paper. http://sites.duke.edu/2014bmp/files/2014/10/Loewen_et_al.pdf.

Mailath, G. and L. Samuelson (2006). *Repeated Games and Reputations: Long-Run Relationships.* Oxford University Press.

McDermott, R. (2002). Experimental methodology in political science. *Political Analysis 10*(4), 325–342.

McGillivray, F. and A. Smith (2004). The impact of leadership turnover on trading relations between states. *International Organization 58*(3), 567–600.

McGillivray, F. and A. Smith (2008). *A Theory of Interstate Relations, Political Institutions, and Leader Change.* Princeton, N.J.: Princeton University Press.

McGillivray, F. and A. C. Stam (2004). Political institutions, coercive diplomacy, and the duration of economic sanctions. *Journal of Conflict Resolution 48*(2), 154–172.

Mercer, J. (1996). *Reputation and International Politics.* Ithaca, NY: Cornell University Press.

Milgrom, P. and J. Roberts (1982). Predation, Reputation and Entry Deterrence. *Journal of Economic Theory 27*, 280–312.

Miller, G. D. (2012). *Reputation and Military Alliances Before the First World War.* Ithaca, NY: Cornell University Press.

Mintz, A. (2004). Foreign policy decision making in familiar and unfamiliar settings. *Journal of Conflict Resolution 48*(1), 91–104.

Office of the Historian, United States Department of State (1956). Foreign relations of the united states, 1955-1957, volume xxiv, soviet union, eastern mediterranean, document 34, memorandum of discussion at the 280th meeting of the national security council, washington, march 22, 1956. https://history.state.gov/historicaldocuments/frus1955-57v24.

Office of the Historian, United States Department of State (1960). Foreign relations of the united states, 1958-1960, volume ix, berlin crisis 1959-1960, germany, austria: Document 81: A memorandum of conversation. https://history.state.gov/historicaldocuments/frus1958-60v09.

O'Neill, B. (1999). *Honor, Symbols, and War.* Ann Arbor, MI: University of Michigan Press.

Pollack, K. M. (2012). A series of unfortunate events: A crisis simulation of a us-iranian confrontation. Technical Report. Available at: http://www.brookings.edu/research/papers/2012/11/us-iranian-confrontation-pollack.

Powell, G. B. and G. D. Whitten (1993). A cross-national analysis of economic voting: taking account of the political context. *American Journal of Political Science 37*(2), 391–414.

Powell, R. (1990). *Nuclear Deterrence Theory: The Search for Credibility.* New York, NY: Cambridge University Press.

Press, D. G. (2005). *Calculating Credibility.* Ithaca, NY: Cornell University Press.

Rand, D. (2012). The promise of mechanical turk. *Journal of Theoretical Biology 299*(21), 172–179.

Renshon, J. (2015). Losing face and sinking costs: Experimental evidence on the judgment of political and military leaders. *International Organization 69*(3), 659–695.

Renshon, J. (2017). *Fighting for Status: Hierarchy and Conflict in World Politics.* Princeton University Press. Book manuscript.

Renshon, J., K. Yarhi-Milo, and J. D. Kertzer (2015). Democratic leaders, resolve and war: Experimental evidence from the knesset. Working Paper. Available at: http://jonathanrenshon.com/Site/CurrentResearch_files/DemocraciesWarCrises-website.pdf.

Sartori, A. E. (2005). *Deterrence by Diplomacy.* Princeton, NJ: Princeton University Press.

Schelling, T. C. (1960). *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

Schelling, T. C. (1966). *Arms and Influence.* New Haven: Yale University Press.

Schultz, K. (2001). Looking for Audience Costs. *The Journal of Conflict Resolution 45*(1), 32–60.

Selten, R. (1978). The chain store paradox. *Theory and Decision 9*(2), 127–159.

Snyder, G. H. and P. Diesing (1977). *Conflict Among Nations: Bargaining, Decision Making, and System Structure in International Crises.* Princeton, N.J.: Princeton University Press.

Solnick, S. L. (1996). The breakdown of hierarchies in the soviet union and china. *World Politics 48*(2), 209–238.

Stewart, F. H. (1994). *Honor.* Chicago, IL: University of Chicago Press.

Tang, S. (2005). Reputation, cult of reputation, and international conflict. *Security Studies 14*(1), 34–62.

Tomz, M. (2007a). Domestic audience costs in international relations: An experimental approach. *International Organization 61*(4), 821–40.

Tomz, M. (2007b). *Reputation and International Cooperation.* Princeton, NJ: Princeton University Press.

Trachtenberg, M. (2012). Audience costs: An historical analysis. *Security Studies 21*(1), 3–42.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453–458.

Walter, B. F. (2006). Building Reputation: Why Governments Fight Some Separatists but not Others. *American Journal of Political Science 50*(2), 313–330.

Weisiger, A. and K. Yarhi-Milo (2015). Revisiting reputation. *International Organization 69*(2), 473–495.

Wolford, S. (2007). New leaders, reputation, and international conflict. *American Journal of Political Science 51*(4), 772–788.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics 13*(1), 75–98.

ONLINE APPENDICES

A Potential Issues in Our Inference

A.1 Realism of Scenarios and Demand Effects

We consider here two issues related to whether simplified scenarios pose a threat to external validity. First, is the concern of the *unrealistic amplification of treatment*: the version of the causal factor presented in the scenario may be implausibly salient – as compared with what people in a real-world setting would experience – and thus lead to an exaggerated estimate of the effect (though in most cases[30] still of the same sign as the real effect). Second, the experiment may present information in an overly abstract and digested format. For example, in our studies we informed the subjects that experts believed that the leader had extensive or limited control over foreign policy but in the real world this information would not be presented as directly, and instead would be part of observers' background knowledge about the country, in part implied by other facts about the country, and in part buried in media that devotes more attention to other aspects of the country and crisis. Accordingly, could it be that our influence treatment was artificially salient and direct, unrealistically cuing the respondent to draw the reputational inference that we found?

Such concerns about the mapping between experimental effects and real-world effects are important. However, we should resist arbitrarily downgrading the informativeness of these results because of these concerns. First, given the difficulty of causal inference in international relations, IR scholars should be (and for the most part, are) primarily concerned with estimating the existence and sign of effects, rather than their magnitude. Concerns about over-estimating the size of our effect are thus far less pressing than the question of whether we have identified an effect in the first place, given the state of our collective knowledge. In fact, a virtue of experiments is that they allow us to isolate and identify effects which would otherwise be hard to detect in noisy real-world data. Second, there are other reasons (low stakes, inattentive and non-expert audience) for believing that our estimates are actually smaller than their real world counterparts. Third, we implemented the

[30] The exceptions being when the causal effect reverses at large doses.

detailed US-Iran scenario precisely to add crucial realism; strikingly we found similar results even though the treatment text was a small portion of the total vignette (the influence treatment, for example, involved 5% of the words of the total vignette).

Are subjects able to infer the goal of the study?

A related concern could be that respondents are able to infer the goal of our study, and that this influences their responses (specifically, by inclining them to respond in ways consistent with our theory). These have been termed "demand effects" (Zizzo, 2010): "changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior."

We believe it is highly unlikely that our influence-specific reputation effects could arise from demand effects, for the reason that the ISR effect is not a main effect, nor even an interaction effect, but a second order interaction effect. This makes it much less likely that the respondent will have inferred what we are asking about. Consider a regular survey that asked people whether they would disapprove of their leader for backing down from a dispute; the respondent can easily infer what is being asked of them, since it is explicitly written. A survey *experiment* is more indirect, since each respondent only reads one side of the comparison. In order to infer the researcher's intended comparison, a respondent reading that the leader backed down would have to anticipate that other respondents read that the leader did not back down. Nevertheless, a sophisticated respondent who knows about survey experiments could potentially infer the goal of the study. An interaction effect — such as the effect of leader turnover on the effect of backing down — is much harder to infer, since the respondent has to anticipate not only the main effects but also that we are interested in how they interact. Further, they would have to anticipate the precise interaction we are looking for; in our design we had many details–including the character of the disputants, the nature of the dispute, the number and identity of those killed on either side, the character of Iran's demands, and the US-Iran balance of capabilities–any of which a respondent could have guessed would have been a causal factor of interest and could be the basis for an interaction effect. Finally, a second order interaction effect is even harder to grasp, as evident by the difficulty in communicating the quantity of interest itself. We are studying: *(the effect of change in perceived influence of the leader on (the effect of leader turnover on (the effect of backing down on perceptions of resolve)))*. If we expressed this in terms of parameters, we have 8 different cells relevant to our key comparison, only a small subset of possible orderings would give us evidence for ISR. To a respondent who may also think capabilities, the number killed, and the nature of Iranian demands were two-level manipulations, we would have 64 cells, making it even less likely that they would perceive the subset of ordering of these cells that corresponds to our hypothesis. Lastly, interaction effects are generally harder to detect than main effects, and 2nd order interaction effects harder to detect still. Thus, to the extent that our manipulation biases our effects upwards, this can be considered as a design compromise in the context of detecting subtle 2nd order interaction effects.

In sum, future researchers should try to pin down plausible estimates of the magnitude of real-world effects, perhaps by designing more realistic and subtle ways of communicating the influence of the foreign design maker. Similarly, we should keep in mind a crucial scope condition: our theory only predicts that reputations should be influence-specific in domains where the observers can perceive the level of influence of the foreign decision maker with sufficient precision. Nevertheless, we are able to learn much from the more achievable challenge of reliably estimating the sign and existence of the effect, as we have done here.

A.2 Samples and Inferences

The theory of ISR is fundamentally rationalist, and makes no particular claims or assumptions about the nature of the actors making reputational inferences. However, it is worth considering two questions concerning how the nature of our samples affects inferences about both the general public and elite leaders: first, does our sample of the general public allow us to make plausible inferences about elites, and second, is our theory of rationalist observers plausible for members of the general public?

On the first issue, there are several points to consider. The first is whether there are theoretical or empirical reasons to suspect that our public sample would be inappropriate for making inferences about elites. On an empirical level, recent studies that have examined differences between elites and the general population have found little to distinguish them. There is thus at least some preliminary evidence that "...convenience samples can be useful for revealing elite-dominated policy preferences," (Hafner-Burton et al., 2014, 845) and that politicians are equally susceptible to "anomalous decision making," tendencies not moderated by either increased consequences or "experience with democratic decision making" (Loewen et al., 2015; see also Linde and Vis, 2016 for a similar study on elites, the public and susceptibility to framing). Most recently, Renshon et al. (2015) find that members of the Israeli Knesset appear to estimate resolve similarly to their counterparts in a representative sample of Israeli citizens.[31]

Of course, these studies constitute only a few data points in what is undoubtedly a complex set of issues. Studies of elites — let alone paired comparisons of elites and citizens — are rare for the simple reason that samples of elites are extremely hard to acquire. And, as a general rule, research programs should not *begin* on elite samples, but rather progress to them over time and after replications have corroborated preliminary findings. McDermott (2002), for example argues that elite samples are too costly most of the time, and that in many cases, concerns about the external validity and generalizability of results from more accessible samples are over-stated or premature.

That elite samples present logistical challenges does not free us from the burden of considering whether and how not having them could impact the inferences we draw. In fact, elite samples are only strictly necessary when elites and the general public differ on dimensions that are *theoretically-relevant* (Renshon, 2015); if research has not advanced far enough to understand what those differences may be, such efforts are likely to be wasted. One concern in this vein is that leaders and the general public might systematically differ in their likelihood of attributing influence to individuals (powerful leaders may, for example, see the world as being influenced mostly by other, powerful individuals rather than structural forces). In that case, our theory would make differential predictions for elites and the mass public, and our sample of the general public would provide an overestimate of the effect of influence on reputational inferences.

[31] Even in many "elite" studies of decision-making, subjects are often far removed from the actual decision makers of primary interest to IR theories. Renshon (2015)., for example, uses political and military leaders drawn from a mid-career training program at Harvard Kennedy School, while Alatas et al. (2009) use Indonesian civil servants and Mintz (2004) uses military officers. These are certainly more elite than college subjects, but still far removed from the dictators, presidents, leaders of the military and foreign ministry, trusted advisors, and generals who are the primary decision makers in most interstate conflicts. This serves as a reminder that the use of quasi-elite subjects, while interesting and helpful, does not obviate the necessity of extrapolating from one population to another.

Of course, there are a nearly infinite number of dimensions on which national decision-makers *might* differ from the general population. For most of these, the key difference that we envision is that elite observers and decision makers should be more rational in their assessments, compared to the public. Elite observers and decision makers are more informed, selected for and more highly trained in rational reasoning and are closer to cultures that encourage rational assessment of foreign policy threats relative to the typical member of the public. Further, the consequences for decision makers are greater if they "get it wrong." All of this suggests that elite observers and decision makers would be better situated and would face powerful incentives to make accurate inferences. Thus, on most dimensions, the public should be less likely to exhibit the rationality of our ISR predictions; if the public is less rational than elites, then we would expect to see weaker evidence for influence-specific reputation and the effects we do see should be interpreted as attenuated relative to the effects that would arise with an elite sample, providing a lower bound on the magnitude of the elite-level effects.

The second broad issue is whether our rationalist theory of reputational inferences might plausibly describe the general public, who might be disinterested and less than knowledgeable about politics. This matters because the public's ability to make these kinds of inferences is crucial to our theory as well as others extant in the literature. For example, one of the main ways in which reputational concerns have been invoked in IR has been through the study of domestic audience costs (Fearon, 1994; Tomz, 2007a). Fearon (1994, 580), for example, argues that it is the precise fact that crises are carried out "in front of political audiences evaluating the skill and performance of the leadership" that allows states beholden to public opinion to send credible signals of resolve. In this influential theory, at least, the mass public is integral.

Here, we can use our results to shed some light on the question of whether the public is apt to make rational inferences. We cannot discount evidence from other sources that the public is short-sighted, or lacks awareness and knowledge about politics. However, if the public is disinterested and not approximately rational then we should not have found evidence for our ISR theory, nor the many other predictions consistent with rational inference, such as the main effects we found. More broadly, while pathologies of human reasoning have famously been identified, it is worth remembering that for most problems, especially those that we confront often or have analogs in our ancestral environment, humans are approximately Bayesian (on human inference as approximately rational, see Holyoak and Cheng (2010).) In particular, inferences about the resolve of individuals and groups is something that is valuable in our daily lives and was valuable in our ancestral environment. Thus, it is plausible that evolution endowed us with faculties sufficiently effective at drawing reasonable inferences about reputation.

B A Model of Influence Specific Reputation

There are two (overlapping) ways of modeling reputation: the first models reputation as an inference about some unobserved characteristic ("type"), the second as an inference about the equilibrium (Mailath and Samuelson, 2006). For tractability we sketch here a simple model of reputation as an inference about type.

B.1 The Stage-Game

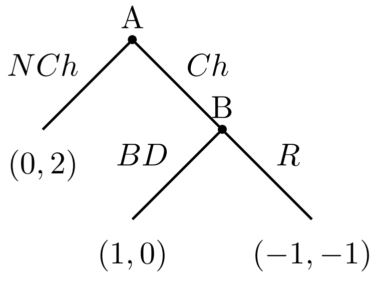


Figure 6: Coercion Stage-Game

Consider the stage-game depicted in Figure 6. This stage-game is a form of sequential move game of chicken and is a variant of the chain store game stage-game (Selten, 1978; Kreps and Wilson, 1982; Milgrom and Roberts, 1982). Fudenberg and Tirole (1995, Ch 9) provide a useful review of these kinds of models.

A would prefer to **Ch**allenge B (to choose Ch) if and only if B will **B**ack **D**own (choose BD). Given that A **Ch**allenges, B would prefer to **B**ack **D**own. B most prefers that A does **N**ot **Ch**allenge B to begin with. In such a game, with complete information and common knowledge of the game, there is a unique subgame perfect equilibrium: A infers that B will **B**ack **D**own if **Ch**allenged; A **Ch**allenges; B **B**acks **D**own.

The problem facing B is that if B could only convince A that B was committed to **R**esisting, then A would be deterred, B would realize his preferred outcome, and B would never have to fight. This characterizes the essence of why being able to endogenously generate commitments can be so beneficial in coercive encounters (Schelling, 1960; 1966).

Now suppose that there is some probability, β, that B is a type of actor who always prefers **R**esisting over **B**acking **D**own. The literature often refers to this type by various adjectives, such as "tough," "crazy," or "extreme". We refer to this type of actor as "intrinsically honorable" because we believe the concept of honor better represents the logic of resistance to coercion.[32] Agents who are not intrinsically honorable types are called "materialist." We can operationalize intrinsic honor by adding a utility cost (of more than 1) to B backing down (BD).

A will then only **Ch**allenge B if B is not too likely to be intrinsically honorable (where ω_A denotes A's strategy) :

$$EU_A(\omega^A = Ch) \geq EU_A(\omega^A = NCh)$$

$$\iff -\beta + (1-\beta) = 1 - 2\beta \geq 0$$

$$\iff \beta \leq 1/2$$

B.2 Reputation Building

Suppose now the stage-game is played twice, with different challengers (Countries in the A role) across each period. We can equivalently think of the game being played twice with the same country A, but that A is completely myopic so that A's discount factor δ_A is 0.

In the final round, B will simply behave according to B's intrinsic honor. Intrinsically honorable types will fight, materialist types will back down.

B.2.1 A in the Final Round

A will challenge so long as the probability of facing an intrinsically honorable B is less than 1/2. Denote this (the probability that B will choose **R**esist in the last round, given that B chose **R**esist in the first round) as $P(a^B_{t+1} = R | a^B_t = R) = \beta_{t+1}$. The subscripts denote time periods, with the last period set to $t+1$, and a^B_t as B's action at time period t. By common knowledge of the game, the equilibrium, and rational updating, this will also be A's belief about the probability of

[32] Note that honor generally contains expectations of being tough towards and "irrational" about the costs of conflict. The primary difference between toughness, irrationality, and honor, regards selection into disputes. A tough agent should be expected to pick more fights than a non-tough agent. A crazy agent should similarly be expected to engage in irrational provocations, as well as other irrational actions. An honorable agent can be as or more circumspect about initiating conflicts; however, conditional on being challenged an honorable agent will stand firm (O'Neill, 1999; Dafoe and Caughey, 2016). Below we discuss more the theoretical resonance between this model and the concept of honor.

facing an intrinsically honorable type given that B fought in round t. Denote the probability that a materialist B will fight in the first round as $p_t = P(\omega_t^{MB} = R)$, where ω_t^{MB} denotes materialist B's strategy in round t. Then by Bayes rule:

$$\beta_{t+1} = P(a_{t+1}^B = R | a_t^B = R) = \frac{\beta}{\beta + (1-\beta)p_t}$$

Then, having seen resistance in the first round, A prefers to **N**ot **Ch**allenge in the last round if

$$EU_A(\omega_{t+1}^A = Ch | a_t^B = R) \geq EU_A(\omega_{t+1}^A = NCh | a_t^B = R)$$

$$\iff \beta_{t+1} = \frac{\beta}{\beta + (1-\beta)p_t} \geq 1/2 \iff \frac{\beta}{1-\beta} \geq p_t \iff \frac{p_t}{1+p_t} \geq \beta$$

This implies that if materialist B never **R**esists ($p_t = 0$), then having seen resistance in the first round A will **N**ot **Ch**allenge in the last round (because A is certain that B is intrinsically honorable). It also implies that if materialist B always resists ($p_t = 1$), then A will only be willing to **Ch**allenge in the last round if there are not too many intrinsically honorable Bs ($\beta \leq \frac{1}{2}$); if $\beta > 1/2$ then A will never challenge. Since we restrict ourselves to the more interesting situation where $\beta < 1/2$, then we see that if all materialist B's **R**esist, then A will not be deterred (which of course means that materialist B's will not want to resist). Thus, in order for B to build a reputation (choosing to resist, leading some A's to be deterred), it must be that not all materialist B's resist (that is, B must be mixing).

B.2.2 B in the First Round

In the first round, a materialist B (denoted MB) has the option of trying to deter A by "behaving honorably": **R**esisting any challenge. So long as the probability of a materialist B behaving honorably is not too large, relative to the proportion of intrinsically honorable B's ($p_t \leq \frac{\beta}{1-\beta}$), then those A's who observe B resisting previous challenges will prefer to **N**ot **C**hallenge in the last round. We can now determine when materialist B will be willing to earn a reputation for being honorable. Denote the probability that A, having seen B resist in the first round, **Ch**allenges in the last round as q_{t+1}. Materialist B is willing to build a reputation for being honorable if

$$EU_{MB}(\omega_t^{MB} = R | a_t^A = Ch) \geq EU_{MB}(\omega_t^{MB} = BD | a_t^A = Ch)$$

$$\iff -1 + \delta_{MB}(q_{t+1} \cdot 0 + (1 - q_{t+1}) \cdot 2) \geq 0$$

$$\iff \frac{2\delta_{MB} - 1}{2\delta_{MB}} \geq q_{t+1}$$

$$\iff \delta_{MB} \geq \frac{1}{2(1 - q_{t+1})}$$

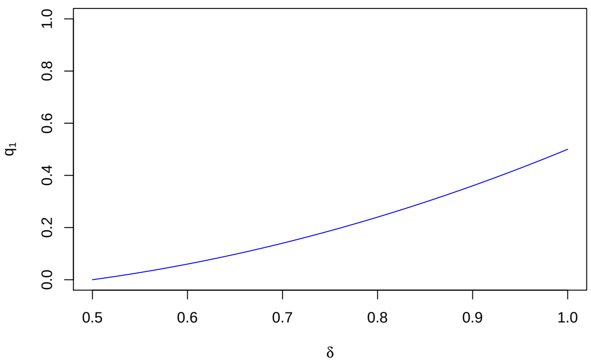


Figure 7: Values of q_{t+1} that make MB indifferent.

For $\delta_{MB} = 1$ the q_{t+1} that makes MB indifferent is 0.5 (there has to be a 50% chance that A can be deterred); for $\delta_B = .75$ it is 0.1875 (there needs to be an 81% chance that A is deterred). The general pattern is plotted in Figure 7. So long as the probability that A will challenge a B who behaved honorably is sufficiently low, relative to B's patience (below the blue line), materialist B will be willing to build a reputation for being honorable.

B.2.3 A in the First Round

A will **Ch**allenge in the first round so long as the probability that B resists is less than a half:

$$EU_A(\omega_t^A = Ch) \geq EU_A(\omega_t^A = NCh)$$

$$\iff \beta + (1 - \beta)p_t \geq 1/2$$

B.2.4 Equilibrium

To summarize, in the last round B will only **R**esist if B is intrinsically honorable. A will be deterred (will be willing to **N**ot **Ch**allenge, having seen B **R**esist before) in the last round if $\frac{\beta}{1-\beta} \geq p_t$.

Materialist B will be willing to **R**esist A in the first round if $\delta_{MB} > \frac{1}{2-2q_{t+1}}$.

For $\delta_{MB} < 1/2$, materialist B is never willing to build a reputation for being honorable, since even if A is deterred with certainty ($q_{t+1} = 0$), the costs to B outweigh the benefits. In this case A will **Ch**allenge in the first round, and in the last round if and only if B backed down in the first round.

Now consider $\delta_{MB} > 1/2$. B may now be willing to build a reputation for being honorable. If A is fully deterred by resistance ($q_{t+1} = 0$), then B will always **R**esist in the first round ($p_t = 1$). But then A will not be deterred (since $\beta < 1/2$), so this cannot be an equilibrium. A cannot be fully deterred, and therefore must mix: $q_{t+1} > 0$. In order for A to accept this, it must be that $p_t = \frac{\beta}{1-\beta}$. Since $0 < \beta < 1/2$, in order for A to mix it must be that some materialist B resist in the first round ($p_t > 0$) and some materialist B do not resist ($p_t < 1$). Therefore, B must also mix, requiring that $q_{t+1} = \frac{2\delta_{MB}-1}{2\delta_{MB}}$. Thus, when $\delta_{MB} > 1/2$, in the last round A will **Ch**allenge a B who had resisted in the first round with probability $\frac{2\delta_{MB}-1}{2\delta_{MB}}$; a materialist B will **R**esist a challenge in the first round with probability $\frac{\beta}{1-\beta}$.

A in the first round will then be willing to **Ch**allenge if and only if $\beta + (1-\beta)p_t \leq 1/2 \iff \beta \leq 1/4$. If $1/4 < \beta < 1/2$, A will **N**ot **Ch**allenge in the first round, but then will **Ch**allenge in the last round, knowing that a materialist B will no longer act honorably. For the purposes of studying reputation, we focus on when $\beta \leq 1/4$.

To summarize, when $\beta < 1/4$ A will **Ch**allenge B; intrinsically honorable B's will **R**esist, materialist B's will **R**esist with probability $p_t = \frac{\beta}{1-\beta}$. If B does not **R**esist then A will **Ch**allenge in the last round. If B does **R**esist, then A will **Ch**allenge with probability $q_{t+1} = \frac{2\delta_B-1}{2\delta_B}$. Only intrinsically honorable B's will **R**esist in the last round.

The above result corresponds to the theoretical distinction between "internal" and "external" honor (Stewart, 1994, 12; O'Neill, 1999, 88-89). Of the $(\beta + (1-\beta)p_t = 2\beta)\%$ of the B's who would build a reputation for being honorable in round t (have external honor), only half of them do this because they are intrinsically honorable (have internal honor). Similarly, this model maps on to a theoretical tension in the theory of honor. Intrinsic honor is most clearly demonstrated when behaving honorably comes at great cost and no one is watching; in our game this occurs in the last round, which is the only time the intrinsically honorable types fully separate from the materialist Bs. In general, however, it is hard to identify intrinsic honor in another person because individuals often have strong reputational incentives to act as if they are intrinsically honorable; in our game materialist Bs (with probability p_t) and intrinsically honorable Bs will choose to behave honorably in all but the last round.

B.2.5 Reputation Hypothesis

We define the change in the conditional probability that B will **R**esist in the last round, depending on whether B **R**esisted in the first round, as

$$\theta = P(a^B_{t+1} = R | a^B_t = R) - P(a^B_{t+1} = R | a^B_t = BD) \tag{1}$$

Assuming rational expectations and updating, θ will also correspond to the change in A's beliefs about whether B is intrinsically honorable. Accordingly, θ operationalizes the effect of B's past actions on observers' perceptions of B's resolve.

If $\theta = 0$, then B cannot acquire a reputation for resolve. In this model $\theta > 0$.[33] If $\delta_{MB} \geq 1/2$ then $\theta = \frac{\beta}{\beta+(1-\beta)p_t} - 0 = \frac{1}{2}$. If $\delta_{MB} < 1/2$, then $\theta = 1$ (since no MB resist, $p_t = 0$, observing resistance implies that B is intrinsically honorable).

Our first hypothesis is that respondents will draw a correctly signed reputational inference, so that previous resolved behavior will lead respondents to increase their beliefs about the probability of future resolved behavior[34]:

$$H_{t+1} : \theta > 0$$

[33] Since θ is only defined for parameter values when A is willing to **Ch**allenge in the first round, this implies that $\beta < 1/4$.

[34] For simplicity, but with some imprecision, we denote respondents' beliefs about θ as θ.

B.3 Reputation with Leaders and Elites

Up to this point, the model has simply formalized how past actions could matter between two unitary actors. Now, we extend the logic to incorporate agent-specific reputations, and, eventually, our theory of influence-specific reputation. To start, suppose that there are two actors in country B: the leader and the elites. The leader is either intrinsically honorable, with probability (β), or materialist, with probability $(1 - \beta)$.[35] Similarly, the elites are either intrinsically honorable, with independent probability (β), or materialist, with probability $(1 - \beta)$. Further, under some (exogenous) circumstances the leader is replaced each round, with a new leader being drawn with independent probability β as intrinsically honorable, and probability $(1-\beta)$ as materialist. Denote when there is a different leader as $L = DL$, and when there is the same leader as $L = SL$.

B.3.1 Low Influence

First consider a country where the leader has low influence ($I = LI$), so that the elites decide B's actions. Then the game is as developed above; we can effectively ignore the existence of the leader. Materialist elites in B will **R**esist in round t with probability p_t. A's in round $t + 1$ who observed B **R**esist a challenge in round t will **Ch**allenge with probability q_{t+1}. The dynamics of leader turnover is irrelevant to the game. We define the country-specific reputation, for a given kind of country, as the effect of past actions under leader-turnover, denoted as:

$$\theta_{CSR,X} = \theta_{DL,X} = P(a^B_{t+1} = R | a^B_t = R, I = X, D = DL) - P(a^B_{t+1} = R | a^B_t = BD, I = X, D = DL) \tag{2}$$

where $X \in \{LI, HI\}$.

We then formalize the hypothesis that country-specific reputations exist as:[36]

$$H_{CSR} : \theta_{DL,LI} > 0$$

[35] Recall that "honorable" in this context simply means a preference for resistance over backing down in disgrace.

[36] Note that this is an easy test for the existence of country-specific reputations since it conditions on the leader having low influence.

B.3.2 High Influence

Now consider a country where the leader has high influence ($I = HI$), so that the leader decides B's actions. A will now make a different calculation about the probability of facing an honorable B in the last round, given that they faced one in the first round: $\beta_{t+1} = P(a^B_{t+1} = R | a^B_t = R)$

When there is leader turnover, and types are independent across time, A learns nothing about B's type in round $t+1$ from B's behavior in round t. Accordingly, B cannot act honorably in round t in order to deter A in round $t+1$. Therefore $P(a^B_{t+1} = R | a^B_t = R, I = HI, D = DL) = \beta$ and $P(a^B_{t+1} = R | a^B_t = BD, I = HI, D = DL) = \beta$, and reputations do not form: $\theta_{DL,HI} = 0$.

Now suppose the same leader is in power in both rounds: $D = SL$. The game will now be equivalent to the basic game, in which reputation building effects are present. $\theta_{SL,HI} > 0$

Putting these together, we can formalize our expectations that there should be leader specific reputation under the high influence condition:

$$H_{LSR} : \theta_{LSR,HI} = \theta_{SL,HI} - \theta_{DL,HI} > 0 \tag{3}$$

B.4 Influence Specific Reputation

Finally, we can define influence specific reputation by seeing if the strength of LSR increases with influence. According to the above model, the effect of past actions under low influence will be the same whether the same leader is present or a different leader, since what matters are the elites' preferences, not the leader's. Thus $\theta_{SL,LI} = \theta_{DL,LI}$. We then have:

$$H_{ISR} : \theta_{LSR,HI} = \theta_{SL,HI} - \theta_{DL,HI} > \theta_{SL,LI} - \theta_{DL,LI} = \theta_{LSR,LI} \tag{4}$$

The above model makes stronger testable predictions than is necessary.[37] The reason for this is that the above model assumes extreme levels of influence, where the leader either has no influence, or complete influence. A more complex model would allow both the leader and elites some influence; this way even under **L**ow **I**nfluence, there would be some leader-specific reputation ($\theta_{LSR,LI} > 0$),

[37] For example, the above model predicts that under Low Influence, the effect of past actions will be the same under a different leader as under the same leader.

and under **H**igh **I**nfluence there would be some country-specific reputation ($\theta_{CSR,HI} > 0$). However, the above ISR hypothesis will still hold.

C Confounding

Contrary a prevailing perception, scenario-based survey experiments can suffer from problems of confounding similar to those that plague observational studies (Dafoe et al., 2015). When manipulation of the words in the vignette change respondent's beliefs about aspects of the scenario in unintended ways, change in the outcome may not be attributable to the causal factor of interest (belief about some specific of the scenario). For example, respondents could be more likely to think that the scenario with *Same Leader* involves an autocracy (where leader tenure can be longer) rather than a democracy (where leader tenure is shorter). We employ a placebo question to evaluate this possibility. Specifically, we ask the respondent about their perceptions of how democratic Country **A** and Iran are, with their answer being expressed as a numerical score between -10 (fully autocratic) to 10 (fully democratic), with example countries at intermediate levels (based on the actual Polity IV scale).

We do, in fact, find evidence of potential confounding. In Study 1, respondents are more likely to think the country is an autocracy when reading the scenario involving *Stood Firm*, *Same Leader*, *High Influence*, and/or *Low Power*. In Study 2, by contrast, these placebo tests are only significant for *Stood Firm* and *Same Leader*, and the magnitudes of the associations are smaller in both. In summary, our placebo tests suggest that the problem of confounding may apply to our studies, particularly Study 1. Any characteristics that are associated with our causal factors of interest (like *Stood Firm* or *Same Leader*) in the minds of the respondents are potential confounds of our design.

However, in this case we are not concerned that our results are confounded, for several reasons. First, we have yet to think of or come across an argument that would connect our causal factors of interest to some other cause of resolve. For example, for regime type to account for our main reputation result (the substantively large θ), respondents would have to think that countries that are slightly more autocratic (about 1.5 points on the Polity scale) are also much more resolved (by at least 30%). The magnitude of such an effect is implausible, since if we extrapolate to a 20 point change in Polity score (from full autocracy to full democracy), the change in resolve would have to be close to 100%. This would, in turn, imply that either full autocracies are always completely resolved or full democracies are always completely unresolved, if not both. Other potential confounds that could account for our results are not obvious to the authors.

Second, Study 2 reduces the imbalance on this placebo question by a large margin, as the design should, and likely reduces imbalance on other unmeasured features of the scenario as well. The result is that not only do the qualitative results remain substantively identical, but the magnitudes of the effects remain similar as well. Nevertheless, we raise this issue because scholars should be aware of the possibility that confounding could drive results in scenario-based survey experiments, such as ours, just as observational studies need to devote attention to discussing and evaluating their control strategies. We offer these placebo results as some evidence of the character of possible confounding. If scholars theorize a plausible confound, extensions of this study could diagnose and control for this confound using the methods described in Dafoe et al. (2015).

version:	**Hyp**	**Iran**	**Hyp**	**Iran**	**Hyp**	**Iran**	**Hyp**	**Iran**
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
Stood Firm	-1.63**	-0.51**						
	(0.249)	(0.152)						
Same Leader			-0.41+	-0.32*				
			(0.251)	(0.153)				
High Influence					-2.19**	-0.23		
					(0.246)	(0.153)		
Power Condition							0.90**	0.05
							(0.152)	(0.153)
Constant	0.892**	-3.553**	0.291	-3.649**	1.171**	-3.690**	0.0845	-3.782**
	(0.175)	(0.107)	(0.179)	(0.108)	(0.173)	(0.108)	(0.125)	(0.108)
N	1804	3177	1804	3177	1804	3177	1804	3177

Standard errors in brackets

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

Table 2: **Placebo Tests to Diagnose Possible Confounding**: Statistically significant positive estimates are cyan and significant negative estimates are red. DV is imputed polity score of country involved in dispute described by scenario. Results are coefficients from OLS models. Table shows estimates from our two surveys: Study 1, which used a hypothetical scenarios and countries (**Hyp**) and Study 2, which focused on Iran (**Iran**).

D Demographic Overview of MTURK Sample

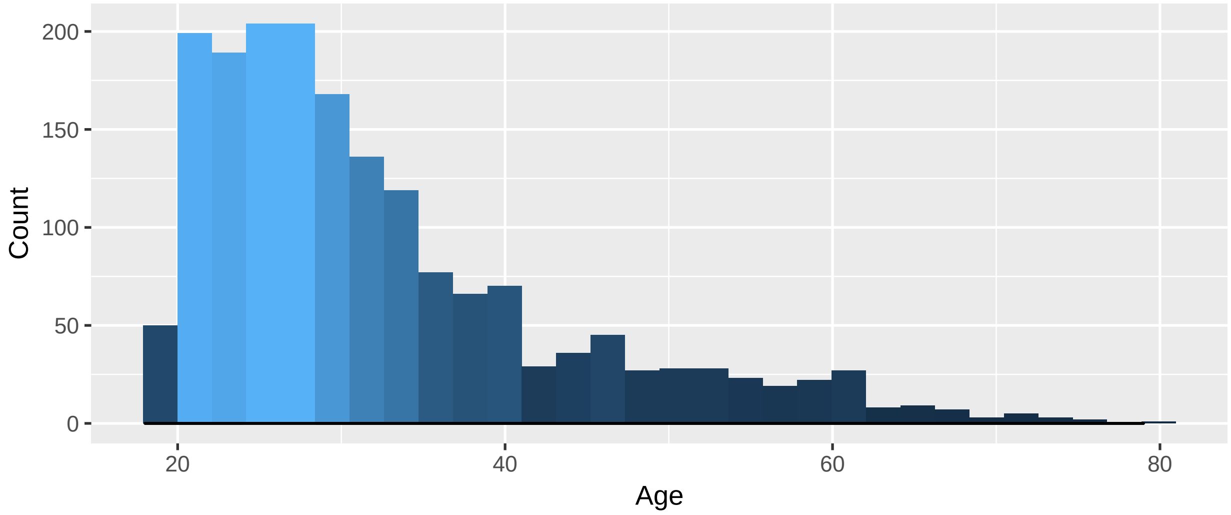


Figure 8: **Age**

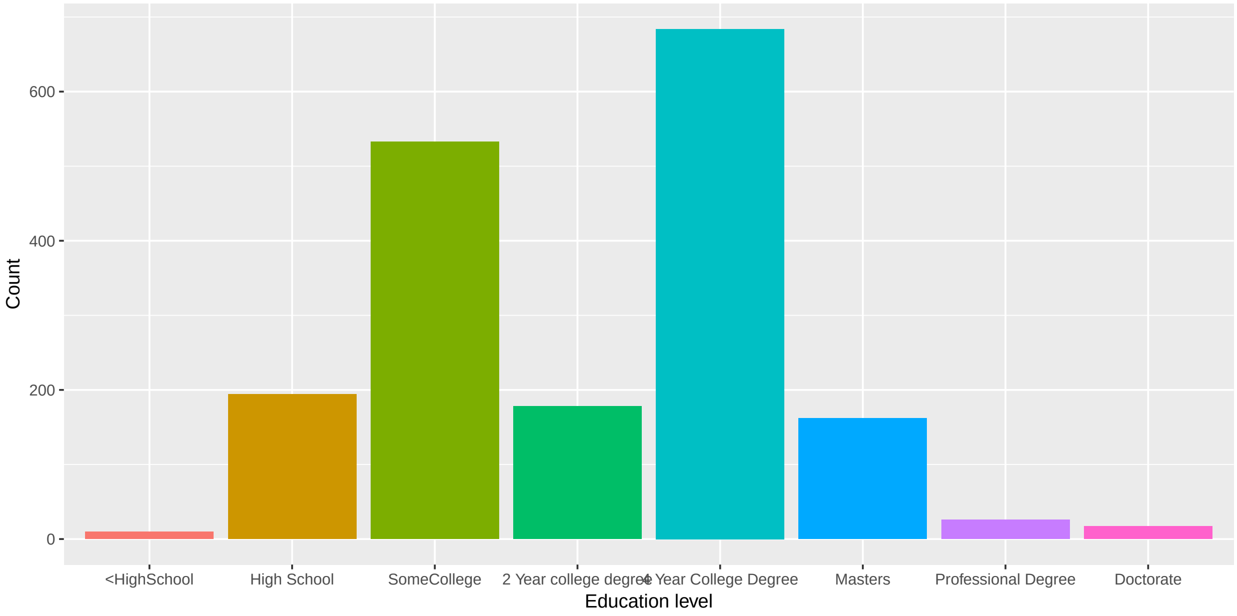


Figure 9: **Education**

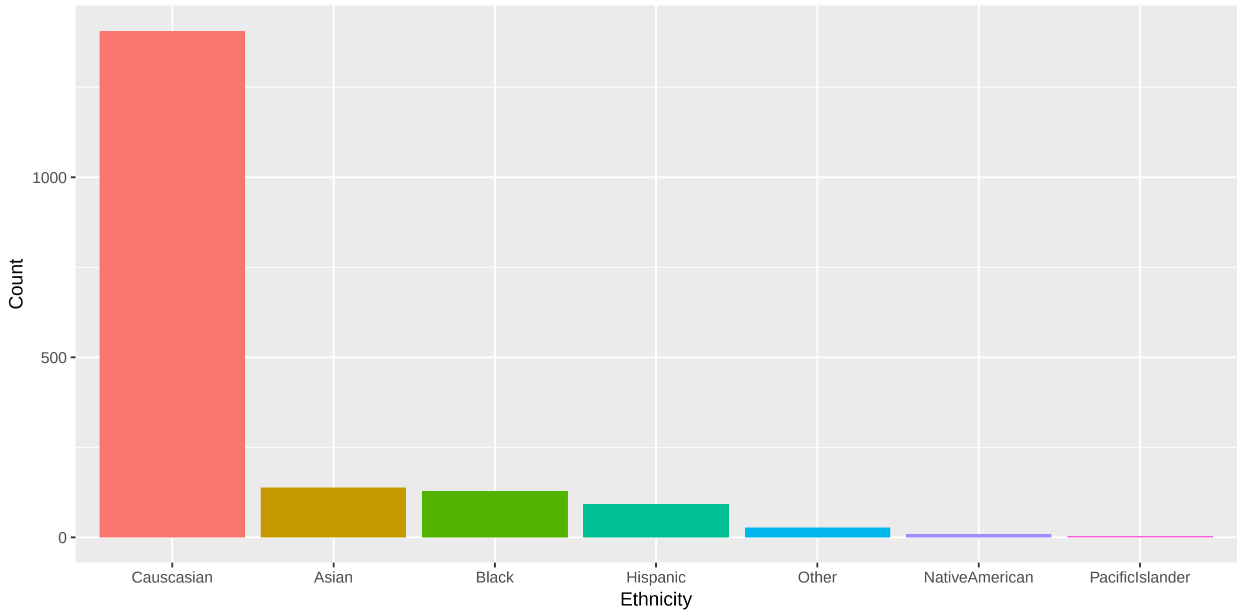


Figure 10: **Ethnicity/Race**

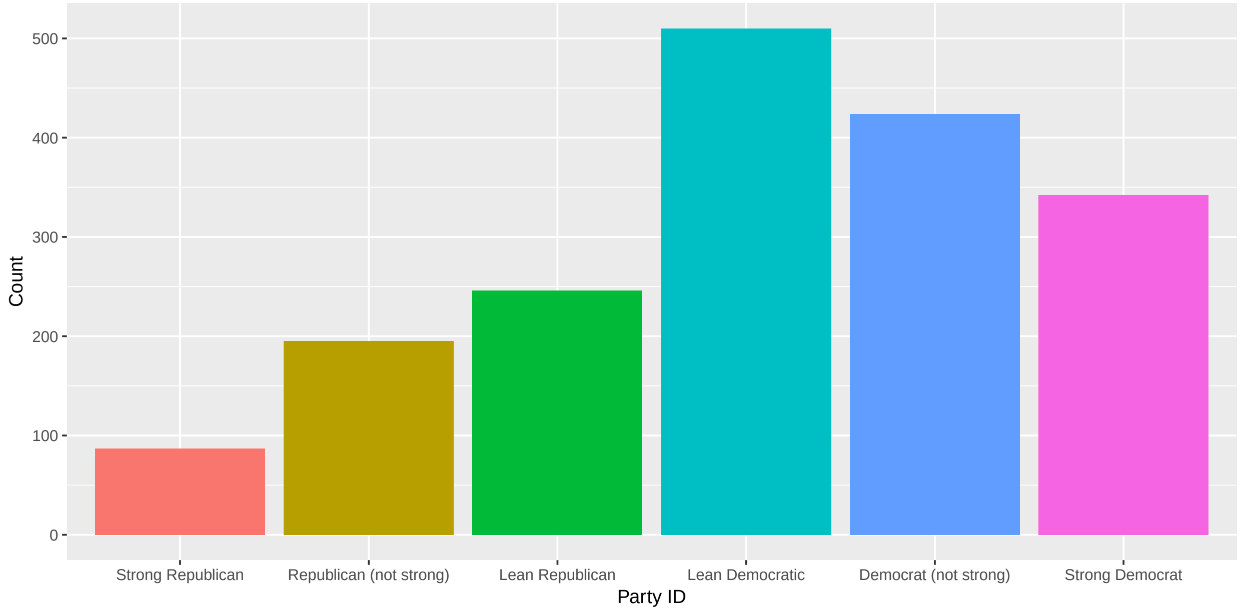


Figure 11: **Party Identification**

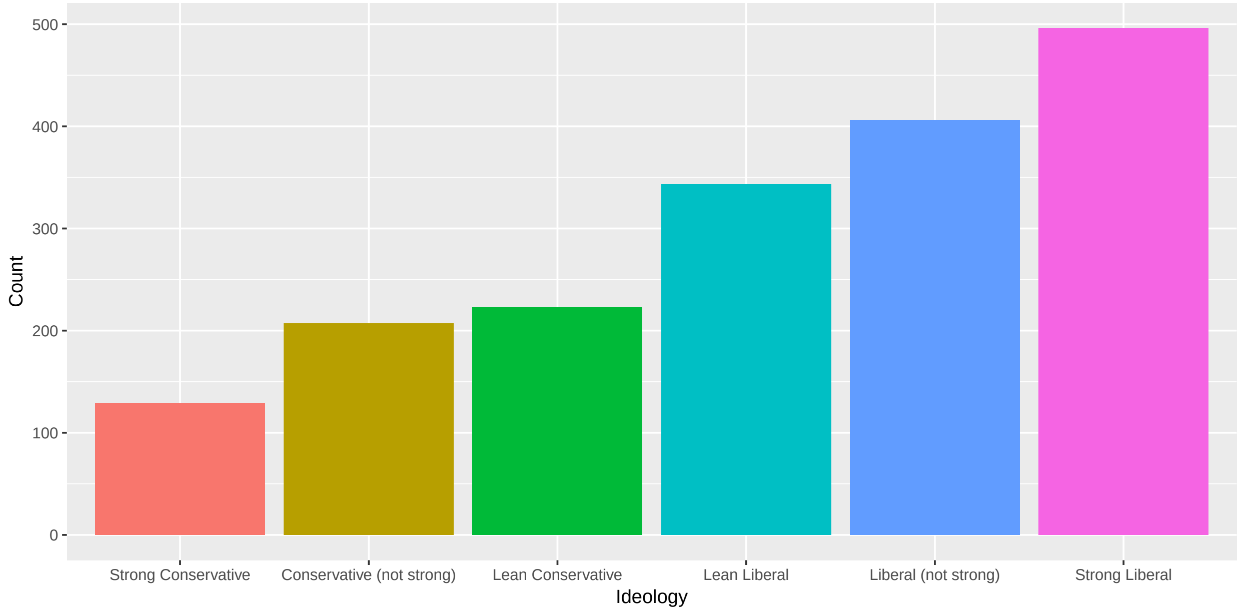


Figure 12: **Ideology**

E Study 1 Vignette

Intro:

Please also remember to read closely and pay attention. During this study you will be asked questions to check your memory and comprehension. You will receive a $0.40 bonus if you answer these accurately, as we expect you should if you read carefully.

The following text will describe a scenario about two countries engaged in a territorial dispute. The countries are labeled Country **A** and Country **B** for purposes of generality.

Please read the scenario carefully and then tell us your beliefs about their likely future behavior.

Initial vignette:

Recent events have led to the flare-up of a serious dispute between two countries — Country A and Country B — over a contested territory. Country A's leader is described by experts as exercising…

complete / very little

control over foreign policy. According to most impartial observers, in the last two (2) international crises between Country A and Country B, Country A…

did not give in to B's demands and did not back down in either crisis / gave in to B's demands and backed down in each crisis

These previous crises occurred under the…

current leader of Country A / **previous** leader of Country A

and under the **current** leader of Country B. The balance of power between Country A and Country B is such that…

Country A is significantly more powerful / Country A is significantly less powerful / Country A and Country B are approximately equal (in terms of military and economic capabilities) than Country B.

Reminder:

To summarize:

- Country A and Country B are involved in a serious dispute over a contested territory
- Country A's leader exercises [complete / very little] control over foreign policy
- in the last two (2) international crises between A and B, Country A [did not give in to B's demands and did not back down in either crisis / gave in to B's demands and backed down in each crisis]
- both of these two previous crises occurred under the [current / previous] leader of Country A and the current leader of Country B
- Country A is significantly [more / less / approximately equal in terms of] powerful (in terms of military and economic capabilities) than Country B

What is your best estimate, given the information available, about whether Country A will back down in this dispute?

NOTE: Answers were scaled from 1 ("Country A is **very likely** to back down [80% to 100% chance]") to 5 ("Country A is **very unlikely** to back down [0% to 20% chance]"), where a "5" represented the greatest estimate of A's resolve in the current crisis.

F Placebo Test Question

1. Now, we would like to ask you about your perceptions of Country A. What is your best estimate of how democratic Country A is, on a scale from -10 to +10, where -10 is fully autocratic and +10 is fully democratic? (Above the slider are some example countries to help you calibrate your answer.)

   Scale was from -10 to +10, with numbers at intervals of 2, and examples at: -10 (North Korea), -6 (China), -2, (Jordan), 2 (Algeria), 6 (Pakistan) and 10 (Canada).

G Manipulation Checks

Can you tell us about the scenario that we just described to you?

1. In terms of control over foreign policy, Country A's leader exercises…
   - no control
   - very little control
   - complete control
2. In the last two international crises between A and B, Country A…
   - gave in to B's demands and backed down in each crisis
   - gave in to some of B's demands, and backed down in only one of the crises
   - did not give in to B's demands and did not back down in either crisis
3. These previous crises occurred…
   - under the **current** leader of Country A and under the **current** leader of Country B
   - under the **current** leader of Country A and under the **previous** leader of Country B
   - under the **previous** leader of Country A and under the **current** leader of Country B
   - under the **previous** leader of Country A and under the **previous** leader of Country B
4. The balance of power between Country A and Country B is such that…
   - Country A is significantly more powerful (in terms of military and economic capabilities) than Country B
   - Country A and Country B are approximately equal in terms of military and economic capabilities
   - Country A is significantly less powerful (in terms of military and economic capabilities) than Country B

H Dispositional Scales & Demographic Information

H.1 Demographic Information

1. How old are you (in years)?

2. What is your gender? [male/ female]

3. What is the highest level of education you have completed? [less than high school/ high school or GED/ some college/ 2-year college degree/ 4-year college degree/ Masters degree/ Doctoral degree/ Professional degree (e.g., JD or MD)]

4. What is your race? [Caucasian/ African-American/ Asian/ Hispanic/ Native American/ Pacific Islander/ Other]

5. What is your combined annual household income? [<30,000/ 30,000-40,000/ 40,000-50,000/ 50,000-60,000/ 60,000-70,000/ 70,000-80,000- 80,000-90,000/ 90,000-100,000/ >100,000]

H.2 Political Ideology & Party Identification

Political Ideology

Generally speaking, would you consider yourself to be a liberal, a conservative, a moderate, or haven't you thought much about this?

- (if Liberal) Do you think of yourself as a **strong** liberal? [yes/no]
- (if Conservative) Do you think of yourself as a **strong** conservative?[yes/no]
- (if Moderate or if haven't thought much about this) Do you think of yourself as more like a liberal or more like a conservative? [liberal/ conservative]

Party ID

Generally speaking, do you think of yourself as Democrat, a Republican, an Independent, or what? [Democrat/ Republican/ Independent/ Other]

- (if Democrat) Would you call yourself a strong Democrat or not a strong Democrat? [Strong Democrat/ Not a strong Democrat]
- (if Republican) Would you call yourself a strong Republican or not a strong Republican? [Strong Republican/ Not a strong Republican]
- (if Independent or if Other) Do you think of yourself as closer to the Democratic Party or the Republican Party? [Closer to the Republican Party/ Closer to the Democratic Party]

H.3 Military Assertiveness

Items 1-8 scaled from 1 (strongly disagree) to 5 (strongly agree). Item 9 scaled from 1 (not very good) to 3 (extremely good) and item 10 scaled from 1 (not at all important) to 3 (very important). Item 2 is reverse coded. Items were presented on one screen, in randomized order.

1. The best way to ensure world peace is through American military strength
2. The use of military force only makes problems worse
3. Rather than simply reacting to our enemies, it's better for us to strike first
4. Generally, the more influence America has on other nations, the better off they are
5. People can be divided into two distinct classes: the weak and the strong
6. The facts on crime, sexual immorality, and the recent public disorders all show that we have to crack down harder on troublemakers if we are going to save our moral standards and preserve law and order
7. Obedience and respect for authority are the most important virtues children should learn
8. Although at times I may not agree with the government, my commitment to the U.S. always remains strong
9. When you see the American flag flying, does it make you feel extremely good, somewhat good, or not very good?

10. How important is military defense spending to you personally? Is it very important, important, or not at all important?

H.4 MTurk Experience

Not including this current study, approximately how many MTURK studies have you participated in…

1. …today?
2. …this week?
3. …in your life?

I Main results (Regression Table)

	θ				*Country & Leader Reputations*				*Influence-Specific Reputations*			
	Hypothetical		Iran		Hypothetical		Iran		Hypothetical		Iran	
	(1a)	(1b)	(1c)	(1d)	(2a)	(2b)	(2c)	(2d)	(3a)	(3b)	(3c)	(3d)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Stood Firm	2.220**	2.186**	1.682**	1.676**	1.761**	1.685**	1.495**	1.480**	1.941**	1.951**	1.550**	1.547**
	(0.0459)	(0.0435)	(0.0344)	(0.0360)	(0.0631)	(0.0592)	(0.0485)	(0.0506)	(0.0800)	(0.0799)	(0.0702)	(0.0735)
Power Condition		0.391**		0.0742*		0.410**		0.0781*	0.411**	0.410**	0.0812*	0.0774*
		(0.0265)		(0.0360)		(0.0254)		(0.0358)	(0.0247)	(0.0247)	(0.0343)	(0.0359)
Age		0.00321		-0.000547		0.00395*		-0.000845		0.00431*		-0.000899
		(0.00206)		(0.00168)		(0.00196)		(0.00167)		(0.00191)		(0.00167)
Male		0.0561		-0.187**		0.0594		-0.181**		0.0334		-0.183**
		(0.0451)		(0.0366)		(0.0431)		(0.0365)		(0.0421)		(0.0365)
Race		-0.0134		0.0115		-0.00377		0.0136		0.00600		0.0129
		(0.0199)		(0.0153)		(0.0190)		(0.0153)		(0.0186)		(0.0153)
Income		-0.00422		-0.00220		-0.00585		-0.00193		-0.00594		-0.00147
		(0.00864)		(0.00704)		(0.00826)		(0.00701)		(0.00805)		(0.00701)
Education		0.0132		0.0195		0.0123		0.0194		0.0105		0.0192
		(0.0168)		(0.0138)		(0.0161)		(0.0137)		(0.0157)		(0.0138)
Republican		0.108		-0.0285		0.0771		-0.0293		0.0891		-0.0287
		(0.0733)		(0.0581)		(0.0701)		(0.0578)		(0.0682)		(0.0578)
Ideology		0.0516*		-0.0333+		0.0426*		-0.0309+		0.0502*		-0.0311+
		(0.0221)		(0.0176)		(0.0211)		(0.0175)		(0.0206)		(0.0175)
Military Assertiveness		0.0434		0.0698		0.0351		0.0804		0.0721		0.0760
		(0.137)		(0.117)		(0.131)		(0.117)		(0.127)		(0.117)
MTurk Experience		-0.00000171*		0.000000160		-0.00000176*		0.000000302		-0.00000173*		0.000000327
		(0.000000845)		(0.00000112)		(0.000000807)		(0.00000111)		(0.000000786)		(0.00000111)
History X Same Leader					0.921**	1.001**	0.372**	0.393**	0.821**	0.819**	0.232*	0.248*
					(0.0888)	(0.0832)	(0.0686)	(0.0717)	(0.114)	(0.114)	(0.0974)	(0.102)
Same Leader					-0.681**	-0.730**	-0.166**	-0.174**	-0.458**	-0.454**	-0.0676	-0.0952
					(0.0626)	(0.0586)	(0.0483)	(0.0504)	(0.0816)	(0.0814)	(0.0680)	(0.0715)
High Influence X Stood Firm X Same Leader									0.368*	0.384*	0.288*	0.288*
									(0.162)	(0.162)	(0.137)	(0.143)
HighInfluence X Same Leader									-0.559**	-0.579**	-0.197*	-0.155
									(0.114)	(0.114)	(0.0966)	(0.101)
High Influence X Stood Firm									-0.520**	-0.540**	-0.103	-0.128
									(0.115)	(0.115)	(0.0970)	(0.101)
High Influence									0.749**	0.763**	0.113+	0.100
									(0.0799)	(0.0801)	(0.0680)	(0.0708)
Constant	2.232**	1.832**	2.148**	2.328**	2.566**	2.211**	2.230**	2.405**	2.242**	1.784**	2.211**	2.356**
	(0.0324)	(0.169)	(0.0242)	(0.141)	(0.0439)	(0.164)	(0.0340)	(0.142)	(0.0560)	(0.166)	(0.0519)	(0.146)
N	1804	1801	3177	2862	1804	1801	3177	2862	1804	1801	3177	2862

Standard errors in brackets

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

Table 3: **Main results.** In the models with *History X Same Leader* interaction, the coefficient on *Stood Firm* represents $\theta_{DL} = \theta_{CSR}$: the effect of reputation when there is a different leader, or the country-specific reputation.

J Study 1 Results by Power Condition

			A has ...					
	all observations		less power		equal power		more power	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Stood Firm	2.220**	2.224**	2.374**	2.369**	2.199**	2.207**	1.976**	1.990**
	(0.0459)	(0.0460)	(0.0716)	(0.0729)	(0.0799)	(0.0799)	(0.0740)	(0.0745)
Age		0.00333		0.00155		0.00701+		0.00222
		(0.00218)		(0.00349)		(0.00369)		(0.00359)
Male		0.0401		-0.00362		0.101		0.0913
		(0.0478)		(0.0742)		(0.0843)		(0.0769)
Race		-0.0216		0.00219		-0.0406		0.00922
		(0.0211)		(0.0337)		(0.0334)		(0.0372)
Income		-0.00132		-0.00246		-0.00772		0.00165
		(0.00915)		(0.0142)		(0.0160)		(0.0149)
Education		0.00541		-0.00878		0.0389		0.00366
		(0.0178)		(0.0287)		(0.0295)		(0.0296)
Republican		0.105		-0.0789		0.106		0.318*
		(0.0776)		(0.119)		(0.137)		(0.126)
Ideology		0.0569*		-0.00232		0.0623		0.101**
		(0.0234)		(0.0362)		(0.0408)		(0.0380)
Military Assertiveness		0.116		0.224		-0.0924		-0.00894
		(0.145)		(0.226)		(0.250)		(0.237)
MTurk Experience		-0.00000168+		-0.00000691		-0.00000138		-0.00000430
		(0.000000894)		(0.00000550)		(0.000000916)		(0.00000437)
Constant	2.232**	1.800**	1.784**	1.740**	2.224**	1.613**	2.759**	2.090**
	(0.0324)	(0.179)	(0.0478)	(0.274)	(0.0582)	(0.316)	(0.0533)	(0.293)
N	1804	1801	610	609	589	589	605	603

Standard errors in brackets

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

Table 4: **Main results, by Power Condition**

K Study 1: Effect of Past Actions (with controls)

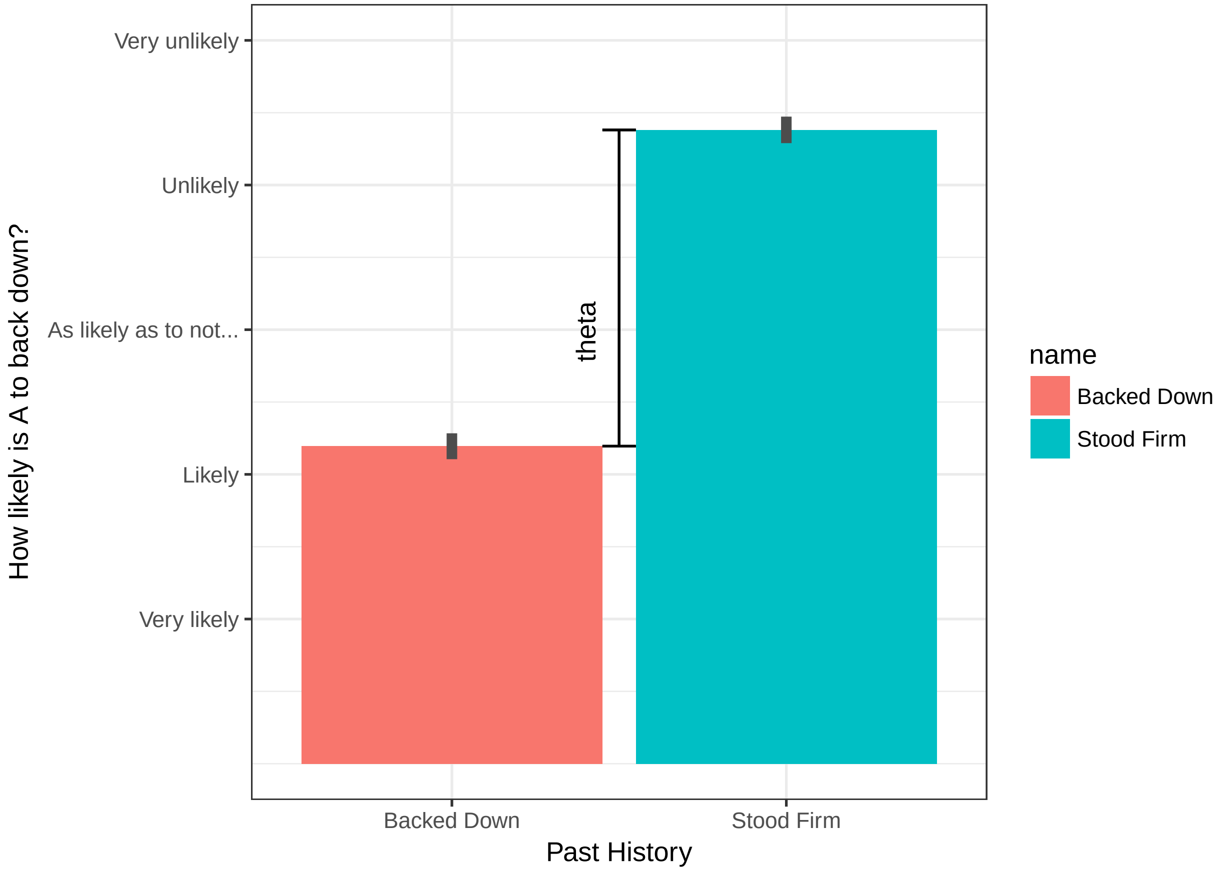


Figure 13: θ, **The effect of past actions on reputations for resolve**: Predicted values generated by holding individual covariates at mean or median using CLARIFY. 95% confidence intervals indicated with vertical lines.

L Study 1: Country-Specific Reputation (with controls)

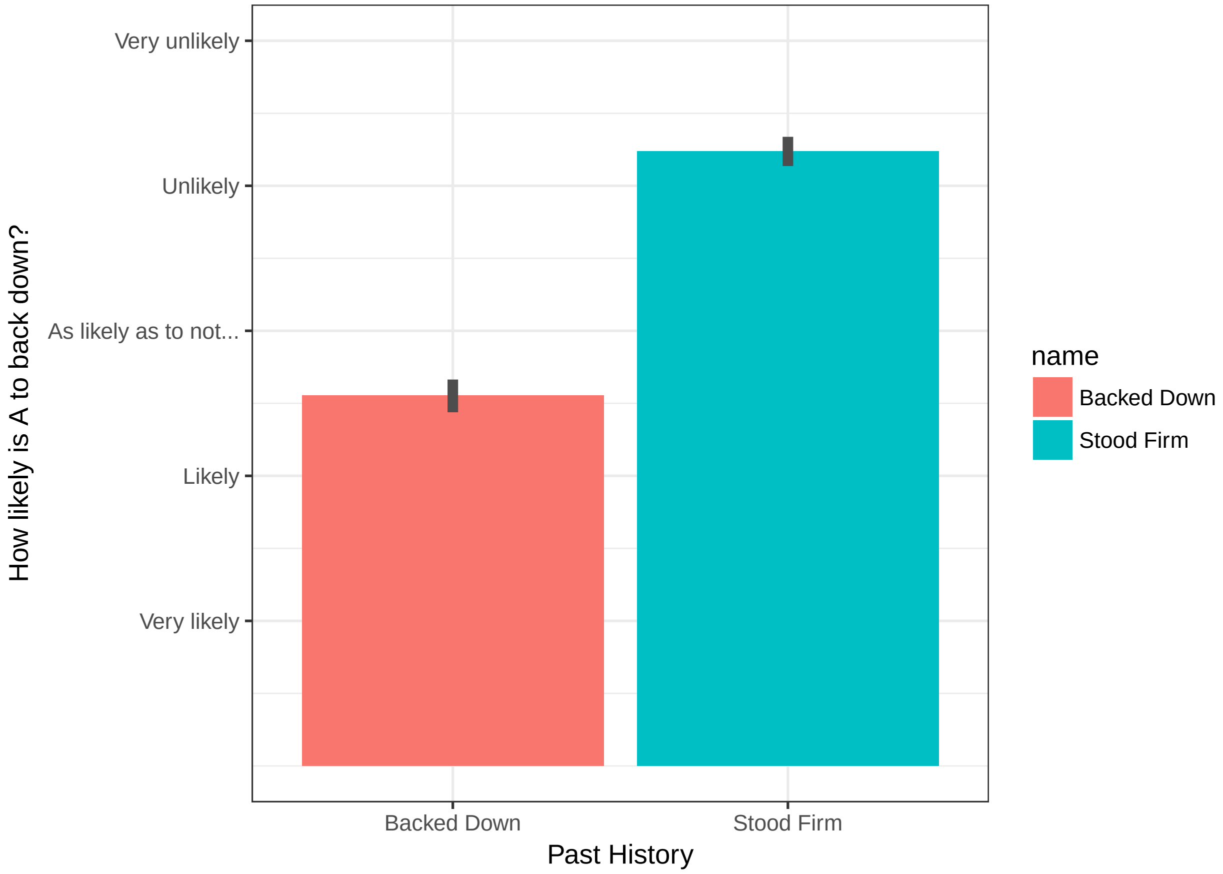


Figure 14: **Country-Specific Reputations**: CSR is equal to θ (the "effect of past actions") when "Same Leader" is set to zero. Predicted values generated by holding individual covariates at mean or median using CLARIFY. 95% confidence intervals indicated with vertical lines.

M Study 2 Vignette: Iran-U.S. Conflict

Intro:

Please also remember to read closely and pay attention. During this study you will be asked questions to check your memory and comprehension. You will receive a $0.40 bonus if you answer these accurately, as we expect you should if you read carefully.

We are going to present you with a hypothetical scenario. Scenarios like this one have happened in the past, and may happen again in the future.

Initial vignette:

It is February 2016. Over the past years, the United States and Iran have been engaged in negotiations about Iran's nuclear program. The U.S. wanted Iran to comply with several United Nations Security Council (UNSC) Resolutions to ensure that Iran doesn't develop nuclear weapons. Iran had not done so, arguing that its nuclear program is peaceful and necessary for its energy security.

Recently, Iran has withdrawn from nuclear talks and is continuing to make progress in enriching uranium. The United States has adopted a policy of "bigger carrots, bigger sticks" toward Iran. Iran has blamed the United States for covert activities, including for a bomb that exploded at the Fordow nuclear enrichment facility, killing 24 Iranians, including six Iranian nuclear scientists.

Following that, terrorists (who the U.S. claims were backed by Iran), retaliated by detonating a bomb at a hotel in Aruba, killing 39 people of various nationalities, including 13 American vacationers and two American nuclear scientists.

The U.S. retaliated by blockading the Strait of Hormuz to Iranian vessels and exports. This blockade will have devastating effects on the Iranian economy. The U.S. has said it will not release the blockade until Iran stops its nuclear weapons program, specifically by complying with UNSC Resolutions.

Experts agree that Iran cannot allow this blockade to continue. Iran must either back-down and comply with UNSC resolutions, or risk war by challenging the blockade.

Iran's president, Hassan Rouhani, is described by experts as exercising...

extensive control over foreign policy; they say that other elites have little influence on Iran's foreign policy/limited control over foreign policy; they say that foreign policy is largely determined by Iranian elites who are independent of Rouhani.

This is not the first time that Iran and the United States have clashed in recent years, though the two previous crises were over much smaller stakes (the release of imprisoned journalists accused of espionage). According to most impartial observers, in the last two international crises between Iran and the United States, Iran...

did not give in to the United States demands and did not back down in either crisis/gave in to the United States' demands and backed down in both crises.

These previous crises occurred under the...

current leader of Iran, Hassan Rouhani/previous leader of Iran, Mahmoud Ahmadinejad...

...and under the current leader of the United States, Barack Obama.

Experts agree that, owing to the current commitment of U.S. forces throughout the world, Iran has...

slightly inferior military capabilities relative to the United States.../significantly less military capability than the United States

...for a conflict over the Strait of Hormuz.

Reminder:

To summarize:

- Iran and the United States are involved in a serious dispute. Nuclear scientists and nationals have been killed on both sides. Iran must either challenge the U.S. blockade of the Strait of Hormuz, or accept the U.S. demands.

- Iran's leader exercises [extensive/limited] control over foreign policy; [other elites in the country have little influence/ it is largely determined by Iranian elites who are independent of Rouhani]

- in the last two (2) international crises between Iran and the United States, Iran [gave in/did not give in] to the demands of the United States and [backed down in both crises/did not back down in either crisis]

- both of these two previous crises occurred under the [previous leader of Iran (Mahmoud Ahmadinejad)/current leader of Iran (Hassan Rouhani)] and the current leader of the United States (Barack Obama)

- owing to the U.S. commitment of forces throughout the world, Iran has [slightly inferior military capability relative to the United States/significantly less military capability than the United States] for a conflict over the Strait of Hormuz

What is your best estimate, given the information available, about whether Iran will back down in this dispute?

NOTE: Answers were scaled from 1 ("Iran is **very likely** to back down [80% to 100% chance]") to 5 ("Iran is **very unlikely** to back down [0% to 20% chance]"), where a "5" represented the greatest estimate of Iran's resolve in the current crisis.