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

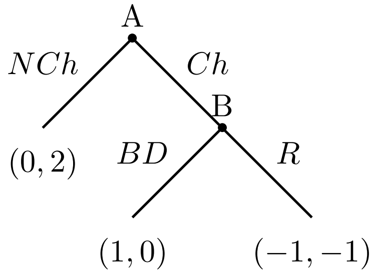

Figure 6: Coercion Stage-Game

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

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

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

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

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

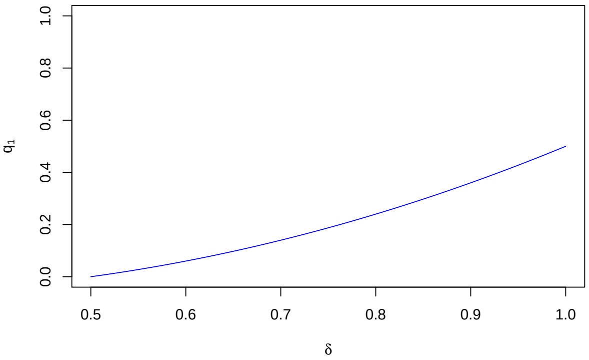

Figure 7: Values of q_{t+1} that make MB indifferent.

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

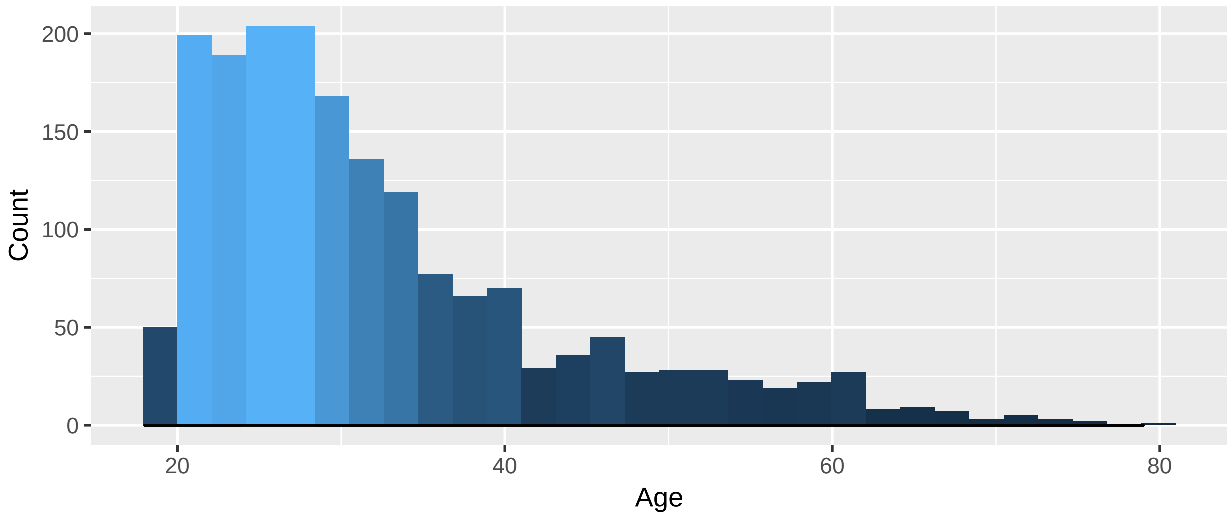

Figure 8: **Age**

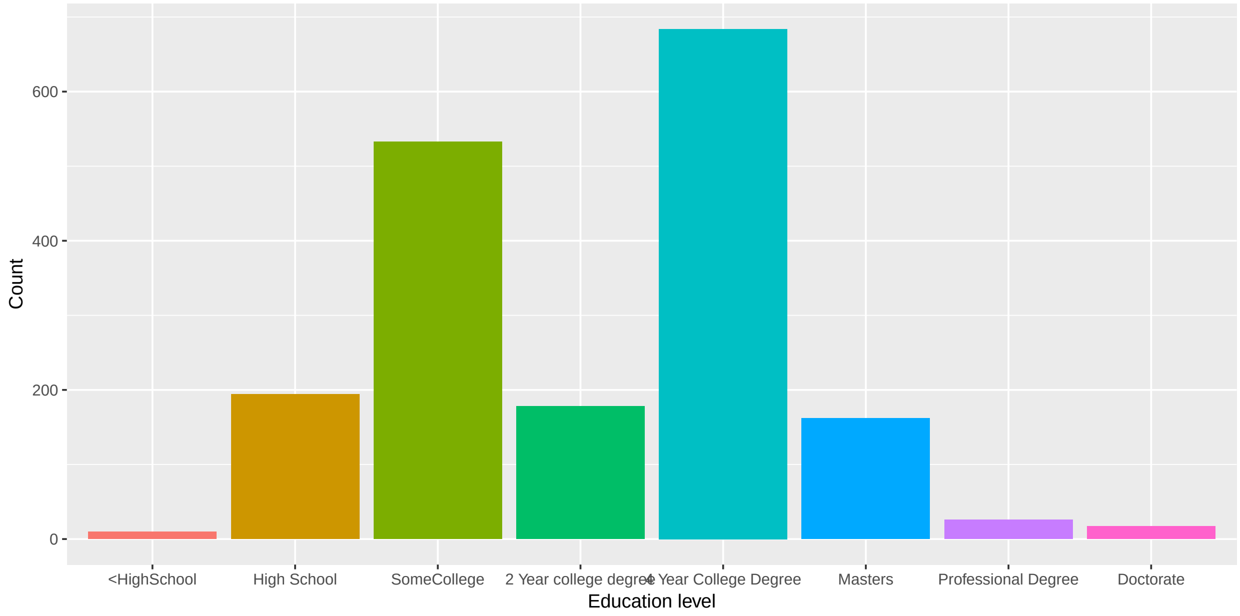

Figure 9: **Education**

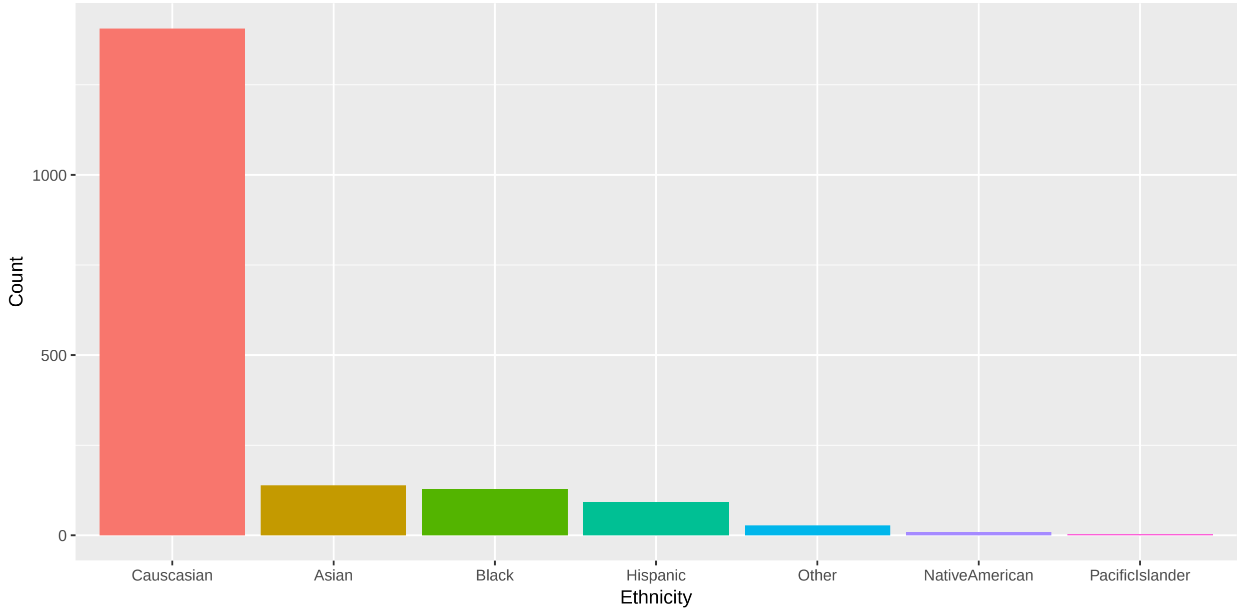

Figure 10: **Ethnicity/Race**

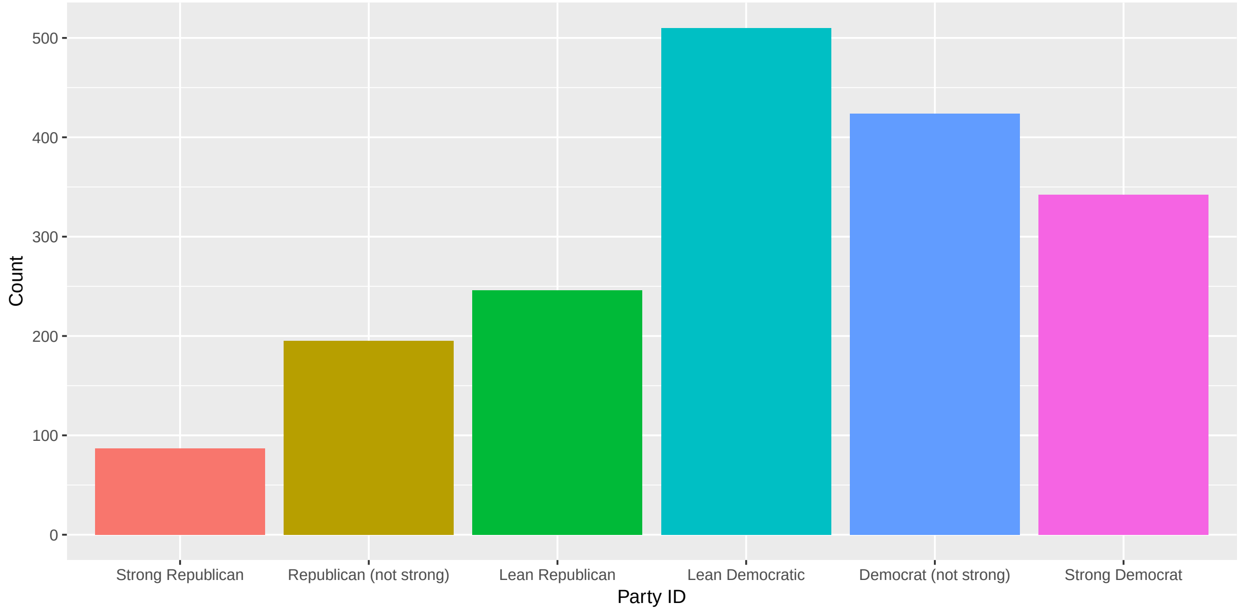

Figure 11: **Party Identification**

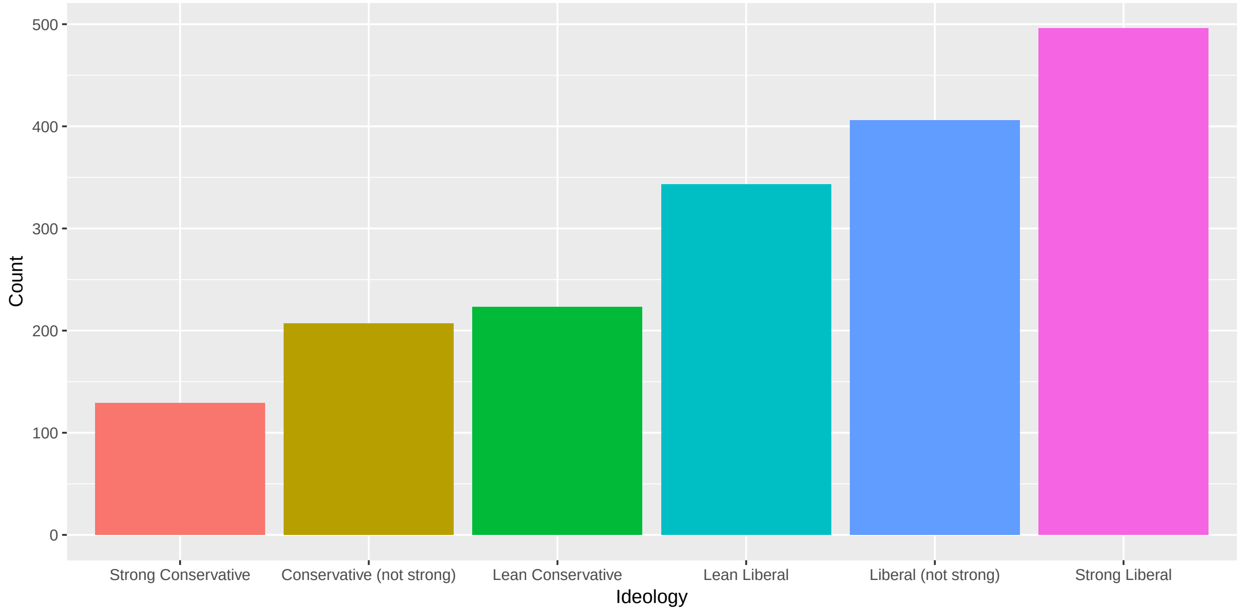

Figure 12: **Ideology**

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

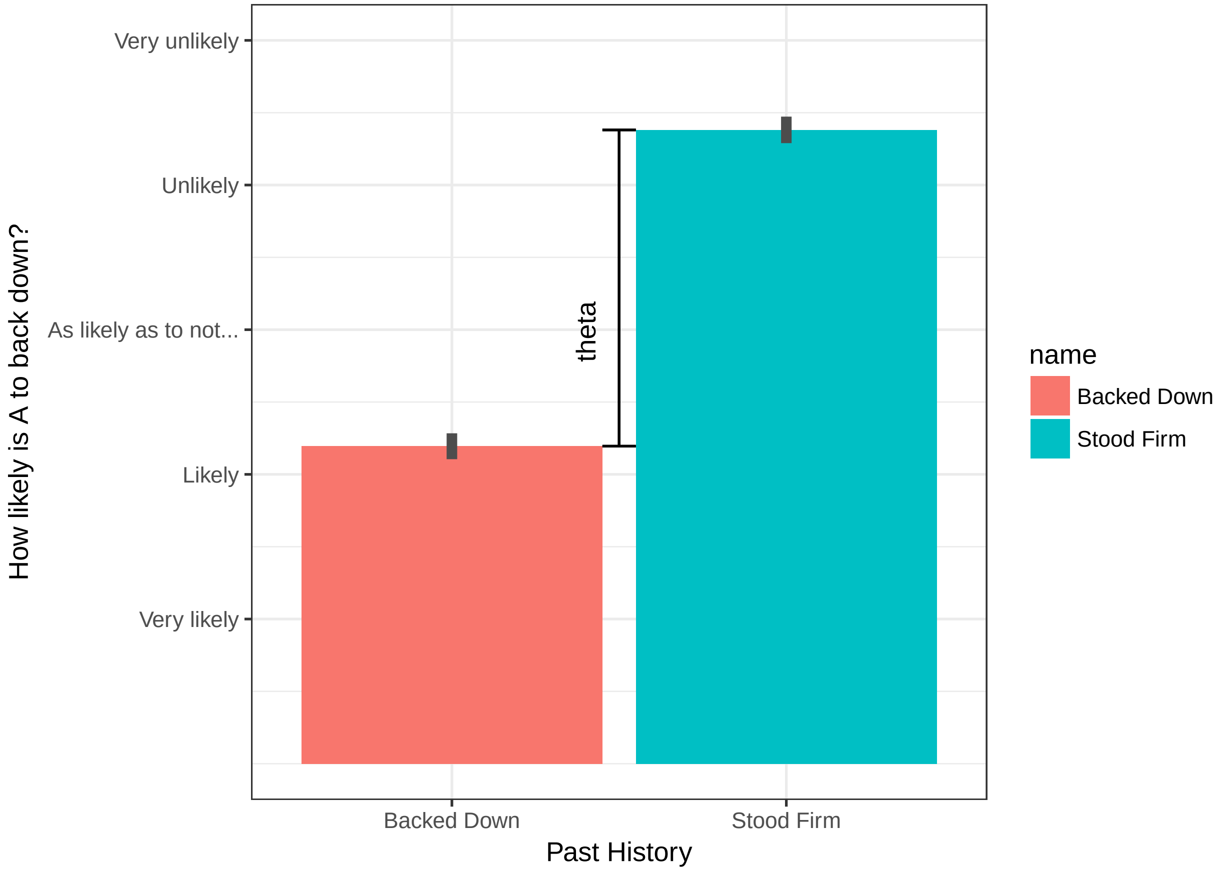

Figure 13: θ, **The effect of past actions on reputations for resolve**: Predicted values generated by holding individual covariates at mean or median using CLARIFY. 95% confidence intervals indicated with vertical lines.

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

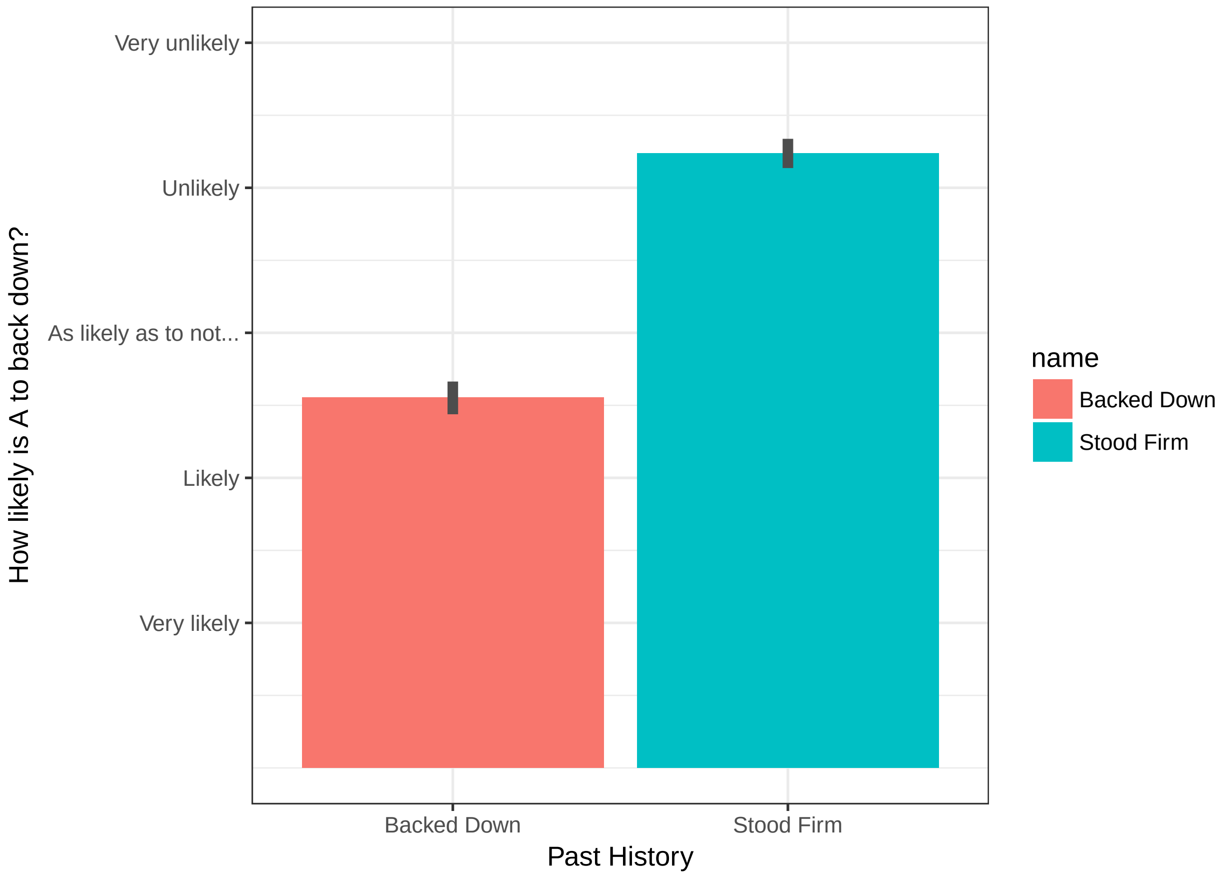

Figure 14: **Country-Specific Reputations**: CSR is equal to θ (the "effect of past actions") when "Same Leader" is set to zero. Predicted values generated by holding individual covariates at mean or median using CLARIFY. 95% confidence intervals indicated with vertical lines.

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.