

How do Observers Assess Resolve?

Joshua D. Kertzer*, Jonathan Renshon† and Keren Yarhi-Milo‡

July 15, 2017

Abstract: Resolve is fundamental to modern theories of international relations, and sits prominently at the center of both theoretical and empirical debates. Yet despite a plethora of theoretical frameworks, IR scholars have struggled with the central question of how we assess the resolve of others, an “ill-structured problem” in which we must adjudicate between an overwhelming number of potential indicators. We innovate on two fronts. Conceptually, we develop an integrative framework that unites (hitherto) disconnected theories, viewing them as a set of heuristics actors use to make sense of information-rich environments. Methodologically, we employ a conjoint experiment that provides empirical traction on a problem that would be extraordinarily difficult to answer using other research designs. We find that ordinary citizens are “intuitive deterrence theorists,” who use heuristics to simplify complex environments and make judgments consistent with a number of frameworks in IR theory. We replicate and extend these results among a sample of elite decision-makers, who converge with the general public on many dimensions, including seeing democracies as less likely to stand firm compared to their autocratic counterparts and perceiving costly signals as important indicators of resolve.

---

*Assistant Professor of Government, Harvard University. Email: jkertzer@gov.harvard.edu. Web: http://people.fas.harvard.edu/~jkertzer/

†Assistant Professor, Department of Political Science, University of Wisconsin-Madison. Email: renshon@wisc.edu. Web: http://jonathanrenshon.net

‡Assistant Professor of Politics and International Affairs, Woodrow Wilson School, Princeton University. Email: kyarhi@princeton.edu. Web: http://scholar.princeton.edu/kyarhi/
Resolve is one of the most central concepts in the study of international politics, used to explain why actors win on the battlefield and prevail at the bargaining table (e.g., Schelling, 1960; Jervis, 1976; Sechser and Fuhrmann, 2017). It is also not directly observable. Indeed, it is precisely because it is unobservable that it is presumed to have such important effects (Kertzer, 2016). Yet, how exactly this process occurs is far from obvious. Intergroup conflicts are characterized by a rich and complex information environment, in which observers can turn to a nearly infinite number of indicators to draw inferences about which actors are more likely to stand firm than others. Assessing resolve in world politics is, therefore fundamentally an “ill-structured problem”: the challenge isn’t one of connecting the dots, but of too many dots to connect (Levy, 1994). Given a large number of varied and countervailing indicators, what ones do observers rely on when assessing resolve?

IR scholars have struggled with this question, and there is still no agreement on what “military capabilities, interests at stake, and past and current actions” lead actors to infer resolve in a given situation (Huth, 1999, 30). And despite a plethora of research in the two decades since that statement was made, we are still no closer to a definitive answer. Political scientists have produced a number of theoretical frameworks to understand inferences about resolve, ranging from the role of past actions (Schelling, 1960, 1966; Jervis, Lebow and Stein, 1985; Mercer, 1996; Weisiger and Yarhi-Milo, 2015), to costly signaling (Morrow, 1994; Fearon, 1997; Fuhrmann and Sechser, 2014), to current capabilities and stakes (Press, 2005), to an ever-growing list of leader attributes (Horowitz and Stam, 2012; Gelpi and Grieco, 2001; Bak and Palmer, 2010). Yet, in providing evidence, we typically focus on a small set of factors in isolation, precluding the development of an overarching conceptual framework to help us understand how these myriad factors work in concert, or provide compelling justification why, from an information-processing perspective, we should expect certain cues to outweigh others.

In this paper, we seek to contribute to this discussion, both conceptually and methodologically. Fundamentally, “which cues do observers use to assess resolve?” is a question about information processing masquerading as a question about IR theory. We therefore draw on a burgeoning body of research in psychology to reconstruct these IR debates and provide an integrative framework that unites these typically disconnected theories, enabling us to view them as sets of heuristics observers can turn to in the face of a computationally intractable information environment. We classify these indicators into two broad classes of indicators — behaviors and characteristics — that individuals may use to assess the likelihood that other actors will back down or stand firm. Methodologically, we test our theoretical framework using a conjoint experimental design ideally suited to answering this question, letting us capture the information-rich nature of international crisis-bargaining environments in a manner that would not be feasible with either case stud-

---

1See also Edelstein (2002); Holmes (2013) and Yarhi-Milo (2014) for the related question of assessment of intentions.
ies, large-N analyses, or even traditional survey experiments. By freeing us from the constraints of focusing on only a small number of treatments and side-stepping the endogeneity and collinearity that threatens our abilities to draw causal inferences using observational data, conjoint experiments allow us to adjudicate between the plethora of competing theoretical frameworks on resolve and credibility that have germinated over the past two decades, testing the observable implications from theoretical frameworks that are rarely investigated together.

Deploying our framework in a conjoint experiment embedded in a survey of 2000 American adults, our results suggest that individuals are intuitive deterrence theorists: ordinary citizens seem to intuitively carry a “folk” version of deterrence theory around in their heads,\(^2\) relying heavily on capabilities, interests, past actions, and costly signals like military mobilization when assessing the resolve of others. For both behaviors and characteristics, our results also challenge the conventional wisdom in IR: in contrast to the significant body of theory on the “democratic advantage” in disputes, our participants see democracies as less likely to stand firm than autocracies, and against recent critiques of reputation in IR, our participants update their assessment of resolve based on past actions. As a robustness check, we also benchmark some of these findings for both behavior and characteristics using conceptual replications on a unique sample of elite decision-makers in the Israeli Knesset, where we find effects of identical direction and strikingly similar magnitude as with the American mass public: our elite decision-makers see democracies as less likely to stand firm compared to their autocratic counterparts, and view costly signals as important indicators of resolve. Our findings thus have implications for both IR theory, public opinion about foreign affairs, and the study of decision-making more generally.

**Heuristics, Cognition and Rationality**

The question of assessing resolve represents a computationally intractable, ill-defined decision problem (Voss and Post, 1988; Levy, 1994; Brutger and Kertzer, 2017), characterized by irreducible uncertainty (Edelstein, 2002). There is a nearly infinite number of indicators observers could turn to in order to assess whether another actor is likely to stand firm or back down in a crisis, many of which often point in different directions. In the case of resolve, this aggregation question is further exacerbated by actors’ incentives to misrepresent (Fearon, 1995). As Yarhi-Milo (2014, 16) notes about assessing intentions more generally, information about resolve “is often complex, ambiguous, and subject to manipulation and deception... cognitive limitations in processing innumerable stimuli, coupled with the need to distinguish usefully and correctly between credible signals and meaningless noise, require the use of some inference strategies or shortcuts.”

\(^2\)For further examinations of the folk theories about international politics ordinary citizens carry around in their heads, see Rathbun (2009); Kertzer and McGraw (2012).
In other words, then, assessing resolve represents the kind of task where we might expect observers to rely on heuristics: “fast and frugal” decision rules that encourage quick and accurate decision-making by focusing on relevant information and setting aside everything else (Gigerenzer and Gaissmaier, 2011, 454-5). Thanks to the pioneering work of Gerd Gigerenzer, Leda Cosmides, John Tooby, and others, the study of heuristics in psychology has undergone a paradigm shift in the past several decades, such that they are now beginning to be understood in a fundamentally different manner than how have traditionally been portrayed in International Relations. Because heuristics were introduced to IR scholars by the “heuristics and biases” literature (e.g., Tversky and Kahneman, 1974), many political scientists understand heuristics as deviations from rationality, a kind of bias or simple-minded cognitive shortcut less likely to manifest itself in high-stakes situations. Press (2005, 6), for example, argues that in high-stakes military matters, “people move beyond the quick and dirty heuristics that serve them so well in mundane matters.” Current thinking on heuristics across psychology and related fields has begun to challenge both of these assumptions.

Whereas political scientists tend to understand rationality as a matter of logical coherence (actors possessing complete and transitive preferences, for example, e.g., Lake and Powell, 1999, 7) psychologists are increasingly understanding rational behavior as a matter of ecological correspondence: not whether actors adhere to the axioms of logic and probability, but whether they behave in ways suited to achieving their objectives in a given environment (Gigerenzer, 2008, 3; Cosmides and Tooby, 1992). One of the counter-intuitive findings of the past few decades of cognitive research has been that for many of the problems we care about, “quick and dirty heuristics” often outperform their more computationally intensive counterparts (Cosmides and Tooby, 1994), a point Axelrod (1984) demonstrates in showing that complex strategies in Prisoner’s Dilemma tournaments are dominated by a simple “tit-for-tat” heuristic that rewards desirable behavior and punishes undesirable ones. This new take on heuristics challenges strongly held assumptions in both rationalist and psychological approaches to IR. It is not that we turn to heuristics because we are “cognitive misers” unable to live up to the lofty optimizing standards of rational ideals, but that since many of the problems we face are either computationally intractable (“NP hard”) or ill-defined, no optimal solutions exist (Simon, 1955; Gigerenzer, 2008; Gigerenzer and Brighton, 2009; Gigerenzer and Gaissmaier, 2011). *Homo economicus* would have never made it out of the Serengeti.

Similarly, although social scientists used to thinking of rational choice as a normative ideal were originally

---

3In this sense, we use heuristics generally to refer to domain-specific decision rules about particular cues (e.g., “if X, then Y”) as in both the public opinion and evolutionary psychology literatures, rather than as content-free algorithms (availability, anchoring, etc.) as in parts of the cognitive psychology literature.

4No consensus is absolute (for example, see the exchange in Kahneman and Tversky 1996 and Gigerenzer 1996), but a broad overview of the literature suggests that whatever disagreements psychologists now have on the utility of heuristics has long since moved beyond IR scholars’ fixation on the “irrationality” of heuristics and biases.

5For critical takes on the divide between psychology and rationality in IR, see McDermott (2004); Mercer (2005); Rathbun, Kertzer and Paradis (2017).
skeptical that heuristics would be widely used in high-stakes situations or by experts with higher level cognitive capacities (e.g., Riker, 1995), we now know that this happens across situations, domains and levels of expertise: doctors use heuristics when making diagnoses where lives are at stake (Green and Mehr, 1997), judges use heuristics when making bail and sentencing decisions (Rachlinski, 2000), and investors use heuristics when making financial decisions (Gigerenzer, 2008, 22). In political science, Loewen et al. (2014) find that national politicians display the same heuristic tendencies as ordinary citizens when given the classic “Asian Disease” experiment from the behavioral decision-making literature. The claim that heuristics bear no impact on the high-stakes world of international affairs may be an article of faith in some quarters, but remains unsupported by evidence.

What indicators do observers use to assess resolve?

What makes the above discussion particularly relevant for our purposes is that it has the potential to speak to a crucial puzzle in International Relations. Given a large number of potential cues, which ones do we rely on the most in assessing the resolve of other actors? Here, we argue that we do not need to generate ideas from scratch, but can rely on extant theories in international relations — of costly signaling, reputation, and so on — as a guide. The critical point is that though we often think of these as theories of international politics, they also contain explicit or implicit hypotheses about the types of cues observers should rely on when predicting whether an actor will stand firm or back down. In order to weave insights together from this diverse array of research, in the discussion below we sweep the conceptual minefield by classifying debates over resolve in IR into two broad classes of explanations, each containing two sub-classes of explanations that advance distinct hypotheses. The first class of explanations emphasizes characteristics of the actor in question, at either the state-level (capabilities, interests, regime type) or leader-level (military experience, time in office, gender, and so on). The second emphasizes behavioral indicators, either in the past (reputation) or in the present (signaling).

Characteristics

State-level characteristics

Observers can calibrate their assessments of an actor’s resolve with reference to the actor’s characteristics at two different levels of analysis. State-level variables may include a state’s level of capabilities, the nature of its interests (generally or in a particular crisis), and its regime type, each of which may impact perceptions of the state’s resolve. For example, a state with greater capabilities may be perceived as more resolute since it
Figure 1: What indicators do we use to assess resolve?

is likely to face lower costs for holding firm compared to a weaker state. Capabilities are often modeled as a source of resolve in game theoretic approaches (e.g., Snyder and Diesing, 1977; Morrow, 1989), and size and strength play important roles in assessments of formidability in evolutionary models (Holbrook and Fessler, 2013). Similarly, a state with high stakes in a situation will likely be perceived as more resolute, such that actors with high levels of interest in the dispute will be more willing to bear the costs they face (Lebow, 1998; Arreguín-Toft, 2001).

Moreover, as the literature on the democratic advantage suggests, democracies tend to outperform autocracies in crises for a variety of reasons, including their ability to be more strategic about the conflicts they enter and display greater initiative on the battlefield (Reiter and Stam, 2002), as well as their advantage at creating audience costs (Fearon, 1994), which we discuss in greater detail below. Based on these various characteristics — capabilities, interests, and regime type — we can generate the following three hypotheses.

*Capabilities and Interests Hypothesis:* States with stronger military capabilities will be perceived as more likely to stand firm than states with weaker military capabilities. States with more interests at stake in a dispute will be perceived as more likely to stand firm than states with fewer interests at stake.

*Regime Type Hypothesis:* Irrespective of who they face in a crisis, democracies will be viewed as more resolute than non-democracies.

**Leader-level characteristics**

A growing body of literature has claimed that particular attributes of leaders affect their inclination to stand firm or back down in crisis, and by implication, others’ assessments of their level of resolve. Along those lines, Horowitz and Stam (2012) find that leaders with prior military service, but not combat experience, are significantly more likely to initiate militarized disputes and wars than other leaders. Gelpi and Grieco (2001)
find that because democracies have a high rate of leadership turnover, democratic leaders have less time in office, and thus less experience compared to their authoritarian counterparts. As a consequence, democracies are more likely to be challenged than autocracies in international crises, since their inexperienced leaders are perceived to be more likely to make concessions. Looking at the effect of leadership turnover on resolve, Wolford (2007) theorizes that observers’ incentive to probe the resolve of new leaders leads the latter to stand firm and develop a reputation for resolve in their early years in office. Building on this, Renshon, Dafoe and Huth (Forthcoming) find that the importance of leader-level characteristics — for inferences about resolve — itself varies according to the amount of influence leaders are perceived to have in a given situation.

Examining the impact of gender on conflict initiation, McIntyre et al. (2007) show that men are more likely to engage in aggressive action than women, and more likely to lose their fights as well, suggesting a relationship between testosterone levels and aggression (see also Horowitz, McDermott and Stam, 2005; Johnson et al., 2006). In addition to biological differences, social expectations may also suggest to observers that female leaders will be less resolute than male leaders in crisis situations (Caprioli and Boyer, 2001). We note, however, that none of these findings is ironclad. Biological differences may push in the opposite direction in some cases (e.g., for older female leaders who have testosterone levels comparable to males, McDermott et al., 2007), and selection effects might result in only extremely tough females coming to power (Anzia and Berry, 2011; Ferreira and Gyourko, 2014).

The insights and findings reported above can help us generate several hypotheses about particular ways leaders’ attributes and experience could affect observers’ assessments of resolve.

 Leaders’ Military Experience Hypothesis: Observers will make predictions about level of resolve on the basis of the military background of leaders. Leaders with military experience are more likely to be perceived as resolved compared to leaders without such background.

 Leader Time-in-Office Hypothesis: Observers will make predictions about level of resolve on the basis of leader’s time in office. Leaders who have been recently elected are more likely to be perceived as resolved compared to more experienced leaders who have been in office for a longer period of time.

 Gender Hypothesis: Observers will make predictions about level of resolve on the basis of leader’s gender. Male leaders are more likely to be perceived as resolute compared to female leaders.
Behavioral Indicators

Past Actions

Behavioral indicators of resolve refer to those actions that actors carry out in order to communicate their intentions to stand firm. A well established literature on reputation in international politics holds that one of the most common ways actors calculate the credibility of current commitments is by looking at past actions (e.g., Copeland, 1997). Both American presidents and political scientists have long assumed that reputation matters in international politics. For much of the Cold War, scholars emphasized the importance of a state’s reputation for resolve as means of deterring future conflicts. The United States’ reputation was seen as one of its most valuable possessions: President Truman justified intervention in Korea on the grounds that a failure to respond “would be an open invitation to new acts of aggression elsewhere,” while Thomas Schelling asserted that the loss of 30,000 men in the resulting inconclusive war was “undoubtedly worth it,” because doing so saved face for the United States and thus positively influenced Soviet expectations of American behavior in future crises (Schelling, 1966, 124-125). Most recently, Weisiger and Yarhi-Milo (2015) have found that states that backed down in the past were more likely to be challenged in subsequent militarized disputes. They report that inferences drawn from past action hold for both democracies and non-democracies and are not reset after leadership turnover.

In contrast to those who believe that reputation matters in international crises, a number of scholars have called into question the effectiveness of past actions in affecting assessments of resolve (Clare and Danilovic, 2012; Danilovic, 2001; Huth and Russett, 1984; Hopf, 1991, 1994; Mercer, 1996; Press, 2005). Jervis (1982), for example, notes that reputational logics lead to a paradox: if actors care about maintaining a reputation for resolve, why should observers assume that an actor who backed down in the past is more likely to back down in the present crisis, rather than more likely to stand firm out of a desire to rebuild their reputation? Others offer critiques that implicate interactions between past actions and other types of indicators. Press (2005), for example, offers a “Current Calculus” hypothesis in which current capabilities and interests override the effects of past behavior. Mercer (1996) offers an equally pessimistic view about reputation, but for different reasons, turning to attribution theory — itself a set of theories about how people process information — to argue that assessments of resolve are affected by the relationship between the observer and the state in question, i.e., whether they are allies or adversaries. Mercer contends that only undesirable behavior by another country — such as when allies back down or adversaries stand firm — elicits dispositional attributions, such that only undesirable behavior can be used to predict future behavior.

The Reputation Hypothesis: Observers will make predictions about level of resolve by looking at the history of the crisis behavior of states. A country with a history of backing down in crisis
will lead observers to predict less resolute behavior, whereas a country with a history of standing firm will lead observers to predict more resolute behavior.

Current Calculus Hypothesis: Observers will make predictions about level of resolve on the basis of calculations of military capabilities and interests at stake rather than on the basis of reputation for resolve.

Attribution Theory Hypothesis: A history of standing firm (backing down) will (will not) lead observers to predict resolute (irresolute) behavior for their adversaries.

A history of standing firm (backing down) will not (will) lead observers to predict resolute (irresolute) behavior for their allies.

Current Actions

The second set of behavioral indicators concern an actor’s current actions, that is, the signaling behavior of leaders during a crisis. The writings of Schelling (1960), Jervis (1970), Fearon (1997) and others (Powell, 2004; Slantchev, 2005) on crisis bargaining feature one common theme: by taking particular costly gestures or actions that raise the risks of escalation, leaders can effectively manipulate others’ assessments of their resolve. Put differently, because resolve is private information and leaders have incentives to pretend to be resolved even when they are not, they need to take actions that would be costly enough to separate them from those unresolved types. Two types of costly signals have received much attention in the literature: sinking costs and tying hands.  

The first type of costly signal of resolve — “sinking costs” — refers to actions, such as mobilizing troops, that are financially or militarily costly _ex ante._ These are actions that are costly to undertake in the near term but do not impact the payoffs associated with future courses of action. The second type of costly signal refers to actions that tie the hands of leaders by creating costs that leaders will suffer _ex post_ if they don’t follow through on their commitment. Following Fearon (1994), a large literature on audience costs has suggested that democratic states are more credibly able to issue threats than non-democratic ones, because of the presence of a domestic audience that will punish leaders when they back down on threats.  

If this argument is true, we would expect the effect of public threats on assessments of resolve would vary

---

6 A similar signaling literature has also developed in biology, explaining the highly ritualized nature of many animal confrontations, which allow actors and observers to assess potential conflict outcomes without actually engaging in costly conflict itself — e.g., Clutton-Brock and Albon (1979) on approaching and parallel walking among red deer, or Ganswindt et al. (2005) on the signaling value of musth in African elephants.

7 See also Fuhrmann and Secheh (2014); Quek (2016), as well as Kertzer and Brutger (2016), who find that even issuing a threat of force can create a sunk cost.

8 Experimental studies on audience costs, on the other hand, have reported mixed results (Tomz, 2007; Levendusky and Horowitz, 2012; Kertzer and Brutger, 2016; Levy et al., 2015).
with regime type: democracies who issue public threats would be perceived as more resolved than their non-democratic counterparts. However, a vibrant strain of recent work in IR has argued that “democratic-autocratic” distinction is either misleading or overstated. Weeks (2008), for example, shows that there is significant variation within autocratic states, with some types able to credibly generate audience costs at rates similar to their democratic counterparts. Relatedly, Downes and Sechser (2012) show that the purported “democratic advantage” is — if it exists at all — far smaller than previously believed.

Costly Signals Hypothesis: Observers will make predictions about level of resolve on the basis of the crisis behavior of the countries involved. Costly signals that sink costs or tie the hands of leaders will lead observers to predict resolute behavior, whereas a country that fails to send such signals will be perceived as less resolved.

Democratic Credibility Hypothesis: Democracies who issue public threats will be perceived as more resolved than non-democratic states who issue public threats.

Research Design

The advantage of the conceptual framework outlined in Figure 1 is twofold. First, it illustrates the extent to which different quadrants of the IR literature have pointed to very different indicators that can be used to infer resolve, each of which offers an observable implication that can be subjected to empirical testing. Second, it offers a loose conceptual ordering to suggest how these different sets of factors are related to one another.

Testing these large number of competing theories against one another, however, poses a number of methodological challenges. Although many of these theories make causal claims, clean counterfactuals are often hard to come by, and many of our factors of interest are endogenous (are current behavior and past actions ever independent of one another?) or collinear (is regime type ever randomly distributed?), which threatens our abilities to draw causal inferences using observational data. In this context, experimental methods can offer important advantages, leveraging random assignment to eliminate confounding variables and explanations. For reasons of statistical power, however, most experiments in IR tend to focus on manipulating only a small handful of factors at a time, and prior experimental work on credibility and reputation has been no exception (Tomz, 2007; Trager and Vavreck, 2011; Davies and Johns, 2013; Tingley and Walter, 2011). This approach has allowed us to uncover the likely influences on assessments about resolve in isolation, but does not let us address the question we’re interested in here: when presented with a plethora of potential indicators of an actor’s resolve, which ones do observers latch on to, and which ones
To address this issue, we use a conjoint experimental design, which is ideally suited for addressing the question we’re interested in (Hainmueller and Hopkins, 2015). In our conjoint experiment, subjects are shown randomly-generated profiles for two countries, A and B, and told that these two countries are engaged in a foreign policy dispute. Subjects are then asked to indicate which country they think is more likely to stand firm. We then repeat the exercise seven more times, each involving a dispute between a different pair of randomly-generated country profiles.

One of the most significant advantages gained through the use of conjoint designs is in statistical power. Conjoint designs free us from the power constraints that limit traditional factorial experiments by making a small number of assumptions, all of which are either guaranteed by the design itself or verifiable empirically, allowing us to pool observations across choice tasks, despite the fact that there are many more possible country profiles than are ever observed in the study. In Appendix §1, we conduct detailed sensitivity analysis to validate these assumptions and demonstrate the robustness of our results, finding, for example, no evidence that respondents behave differently over time as they become familiar with the study (i.e., “demand effects”). An additional advantage is that paired conjoint designs similar to the one utilized here seem to perform remarkably well in reproducing behavioral data obtained in actual voting and choice contexts (Hainmueller, Hangartner and Yamamoto, 2015). The end result of conjoint experiments is the ability to estimate the effect of each treatment — in this setup, referred to as the Average Marginal Component Effect (AMCE) — with a relatively small number of subjects (N = 2000). The AMCE represents the average difference in the probability of being seen as the country more likely to stand firm when comparing two different attribute levels — e.g., a democracy versus a dictatorship — where the average is taken over all possible combinations of other country attributes (Hainmueller and Hopkins, 2015).

Sample

The study described below was fielded on 2,009 American adults recruited via Amazon Mechanical Turk (MTurk) in January 2015, a widely used resource in experimental social science; we elaborate on our use of the sample in detail in Appendix §4. In this sense, our main experimental results speak to how

---

9 The first assumption is that the potential outcomes remain stable across periods; i.e., a subject should pick a given country (conditional on its and the other country’s attributes) regardless of what countries or choices they had seen previously. Second, we assume no profile-order effects; i.e., we assume the choice would be the same regardless of the order in which the two countries are presented. The final assumption is that the attributes of each profile are randomly generated. Those assumptions are discussed in considerable detail in Hainmueller, Hopkins and Yamamoto (2014). For examples of recent conjoint experiments in IR, see Ballard-Rosa, Martin and Scheve (2017); Huff and Kertzer (Forthcoming).

10 Although nationally diverse, our sample is not representative of the American population as a whole, so to preclude against biased estimates of our treatment effects, in Appendix §4 we follow Huff and Kertzer (Forthcoming) in presenting supplementary results where we employ entropy balancing to reweight the data to population parameters, where we show our results do not significantly differ regardless of whether weights are used.
members of the mass public assess resolve, a decision we make for three reasons. First, what the public thinks has significant ramifications for theories of international relations: the empirical record shows that American leaders carefully monitor public opinion in making decisions about the use of force, intervention, retrenchment, nuclear deployment, and so on. From tempering President Eisenhower’s approach to the Taiwan Crisis of 1958 (Eliades, 1993), to compelling President McKinley’s involvement in the Spanish-American War (May, 1961), to reversing President Clinton’s Somalia intervention in 1993 (Foyle, 1999, 201-229), public opinion frequently shapes American foreign policy during crises. Whether the channel of influence is direct (e.g., when leaders pay attention to the polls - see Tomz, Weeks and Yarhi-Milo 2017 for direct experimental evidence) or indirect (such as through elected members of Congress, as in Gelpi and Grieco, 2015), there is much to be learned from an investigation of the manner in which ordinary citizens evaluate and draw inferences about resolve in international affairs (Kertzer, 2016, 50-51).

Second, the phenomenon we explore here is bigger than just IR: evolutionary theorists argue that selection pressures have hard-wired humans and other animals with cognitive and behavioral mechanisms to enable us to quickly and accurately draw inferences about others’ resolve, or what anthropologists and evolutionary biologists call “resource-holding potential” or “formidability” (Maynard Smith, 1974; Parker, 1974; Parker and Rubenstein, 1981; Sell, Tooby and Cosmides, 2009; for an application of this coalitional psychology to IR, see Lopez, McDermott and Petersen, 2011). The question of how we employ these adaptive mechanisms and weight multiple factors into a single summary representation to assess formidability is thus far from a question that solely applies to high-ranking intelligence officers, as seen by the range of studies that have tested this same question in a wide-range of animals, including five-year old children, for example (e.g., Pietraszewski and Shaw, 2015). Given that we know how toads, speckled wood butterflies, red deer, and African elephants assess resolve (Davies, 1978; Clutton-Brock and Albon, 1979; Ganswindt et al., 2005), it seems germane for us to ask the same question about ordinary humans.

The last point is methodological in nature. One of the reasons for the persistent disagreements in debates about assessments of resolve is that scholars confront a cornucopia of competing theories, but relatively little data to help us adjudicate between them. Although many of these theories make causal claims, clean counterfactuals are often hard to come by, and many of our factors of interest are endogenous or collinear. In this context, experimental methods can offer important advantages in isolating and manipulating causal features of interest. Importantly, though, if we want to harness the advantages of experimental methods, we must — even if we are eventually concerned with the inferences of leaders — by necessity begin with studies on ordinary citizens. Moving on to elite samples makes sense only after replications and extensions have increased our confidence in a given research program and narrowed down the plausible candidates for experimentation to a number of factors feasible for study in the extremely small samples that characterize
elite experimentation. Later on in the paper, we do just that, benchmarking our mass public results with those from elite decision-makers, where we find striking similarities between the two samples that both reassures us about our findings, and raises interesting theoretical questions in its own right.

Experimental design

After the consent process, subjects saw an introductory screen which told participants they would be presented with information about a series of foreign policy disputes over contested territories; in each case, they would be presented with information about a pair of countries involved in a dispute, and asked to make predictions about what they think would occur. Subjects were then presented with the first scenario, which randomly generated attributes for each of the countries in the dispute, and asked participants to indicate which country they saw as being more likely to stand firm. The full list of treatments is presented in Table 1, and a sample scenario is depicted in Table 2. Participants were then presented with seven additional randomly-generated conflict scenarios, such that each respondent performed the task eight times in total.

Either before or after the main study, participants answered a series of demographic questions, including gender, age, education, Party ID, ideology, interest in politics and international affairs.

Two points are worth stressing here. First, our dependent variable is an assessment of resolve rather than “credibility”, a narrower concept that refers only to scenarios where explicit threats or promises are made. In our research design, we look at whether democracies are able to threaten more credibly, but we also examine the broader question of whether democracies are seen as more resolved in general. Second, our list of treatments (Table 1), large as it may be at 15 factors with 27 levels (equivalent to a $2 \times 3 \times 2 \times 2 \times 3 \times 2 \times 2 \times 2 \times 3$ factorial design, with a total of 10,368 cells), is obviously not exhaustive. There are myriad additional state- and leader-level characteristics and behaviors that might be studied in future work beyond those we focus on here, from cultural variables (e.g., Nisbett and Cohen, 1996) to leader age (e.g., Horowitz, McDermott and Stam, 2005). Our aim was to manipulate fifteen factors about which IR scholars have offered a variety of — often contradictory — theoretical expectations, pitting them against one another in a manner that would be difficult with observational data and traditional factorial experiments.

---

11 See Appendix §3 for the full text.
12 Importantly, all of the crises participants are presented with concern territorial disputes, in order to hold the meaning of “standing firm” constant. Future research should examine assessments of resolve in other types of disputes. See Appendix §1.1 for randomization constraints.
13 Demographic questions are reproduced in Appendix §2.
14 As the above notation illustrates, a case study design disentangling these factors would require 10368 cases. On parallels between experimental and case study research design, see Gerring and McDermott 2007.
<table>
<thead>
<tr>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Country-level</strong></td>
</tr>
<tr>
<td>(A) Military capabilities</td>
</tr>
<tr>
<td>(1) ... has a very powerful military</td>
</tr>
<tr>
<td>(2) ... does not have a very powerful military</td>
</tr>
<tr>
<td>(B) Interests/stakes</td>
</tr>
<tr>
<td>(1) ... high</td>
</tr>
<tr>
<td>(1) ... low</td>
</tr>
<tr>
<td>(C) Regime type</td>
</tr>
<tr>
<td>(1) ... a democracy</td>
</tr>
<tr>
<td>(2) ... a dictatorship</td>
</tr>
<tr>
<td>(3) ... in between a democracy and a dictatorship</td>
</tr>
<tr>
<td>(D) Foreign relations</td>
</tr>
<tr>
<td>(1) ... the United States</td>
</tr>
<tr>
<td>(2) ... an ally of the United States</td>
</tr>
<tr>
<td>(3) ... an adversary of the United States</td>
</tr>
<tr>
<td><strong>Leader-level</strong></td>
</tr>
<tr>
<td>(E) Time in office</td>
</tr>
<tr>
<td>(1) ... recently took office</td>
</tr>
<tr>
<td>(2) ... has been in power for many years</td>
</tr>
<tr>
<td>(F) Gender</td>
</tr>
<tr>
<td>(1) He</td>
</tr>
<tr>
<td>(2) She</td>
</tr>
<tr>
<td>(G) Military experience</td>
</tr>
<tr>
<td>(2) has served in the military briefly</td>
</tr>
<tr>
<td>(3) had a long career in the military</td>
</tr>
<tr>
<td><strong>Behavior</strong></td>
</tr>
<tr>
<td>(H) Initiator</td>
</tr>
<tr>
<td>(2) it initiated the crisis</td>
</tr>
<tr>
<td>(I) Identity of other state in previous dispute</td>
</tr>
<tr>
<td>(2) adversary of the United States</td>
</tr>
<tr>
<td>(J) Outcome of previous dispute</td>
</tr>
<tr>
<td>(2) the country ultimately backed down</td>
</tr>
<tr>
<td>(K) Leadership change</td>
</tr>
<tr>
<td>(1) ... led by a different leader than the one in the current dispute.</td>
</tr>
<tr>
<td>(2) ... led by the same leader as the one in the current dispute</td>
</tr>
<tr>
<td><strong>Current behavior</strong></td>
</tr>
<tr>
<td>(L) Costly signals</td>
</tr>
<tr>
<td>(1) ... has yet to make any statements or carry out any actions</td>
</tr>
<tr>
<td>(2) ... has mobilized troops</td>
</tr>
<tr>
<td>(3) ... has made a public threat that they will use force if the other country does not back down</td>
</tr>
</tbody>
</table>

Table 1: Conjoint Study Treatments: Treatment categories are denoted by letters (A-L), while gray blocks denote clusters of treatments that are displayed together in order to remain understandable. All other items are displayed in a random order. Though displayed together, all treatments in gray clusters are manipulated independently save for the constraints imposed on randomization if the country is designated as being the United States (described in detail in main text).
In disputes like these, countries either back down or stand firm.
If you had to choose between them, which of the two countries is more likely to stand firm in the current dispute?

- [ ] Country A
- [ ] Country B

Given the information available, what is your best estimate about whether Country A will stand firm in this dispute, ranging from 0% to 100%? [slider from 0-100]

Given the information available, what is your best estimate about whether Country B will stand firm in this dispute, ranging from 0% to 100%? [slider from 0-100]

Table 2: Sample Conjoint Choice

<table>
<thead>
<tr>
<th></th>
<th>Country A</th>
<th>Country B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>The country is a democracy</td>
<td>The country is a democracy</td>
</tr>
<tr>
<td>Interests in the dispute</td>
<td>Experts describe the country's stakes in the dispute as high.</td>
<td>Experts describe the country's stakes in the dispute as high.</td>
</tr>
<tr>
<td>Leader background</td>
<td>The leader recently took office; he has served in the military briefly.</td>
<td>The leader recently took office; she had a long career in the military.</td>
</tr>
<tr>
<td>Foreign relations</td>
<td>The country is an ally of the United States.</td>
<td>The country is an adversary of the United States.</td>
</tr>
<tr>
<td>Previous behavior in international disputes</td>
<td>The last time this country was involved in an international dispute, it initiated the crisis by issuing a public threat to use force against an adversary of the United States, but ultimately backed down. At the time, the country was led by a different leader than the one in the current dispute.</td>
<td>The last time this country was involved in an international dispute, it initiated the crisis by issuing a public threat to use force against an adversary of the United States, and stood firm throughout the crisis. At the time, the country was led by a different leader than the one in the current dispute.</td>
</tr>
<tr>
<td>Current behavior</td>
<td>In the current crisis, the country has yet to make any statements or carry out any actions.</td>
<td>In the current crisis, the country has made a public threat that they will use force if the other country does not back down.</td>
</tr>
<tr>
<td>Military Capabilities</td>
<td>The country does not have a very powerful military</td>
<td>The country has a very powerful military</td>
</tr>
</tbody>
</table>
Results and Discussion

For purposes of simplicity, in the discussions below we focus only on those crises where participants were either allies of the United States or adversaries of the United States, but not the United States itself. Two rationales drive this decision. Methodologically, estimating the resolve of our direct opponents triggers strong biases (e.g., positive illusions) that we are able to sidestep by focusing on observers. Substantively, the results shown here represent foreign policy disputes where respondents are third-party observers rather than affiliated with the actors themselves, reflecting an important class of crises — the Suez Crisis, the Iran-Iraq War, recurring Indo-Pakistani and Arab-Israeli crises, and so on — in which Americans are onlookers rather than immediate participants. The quantities of interest are calculated from 8,090 choice tasks made by 1,995 participants between 16,180 country profiles.

We begin our analysis by estimating the Average Marginal Component Effects (AMCEs), presented with 95% clustered bootstrapped confidence intervals in Figure 2. These estimates tell us the percentage change in the perceived likelihood that an actor with a particular attribute will be seen as being more likely to stand firm in a foreign policy dispute. Since, with the exception of the past behavior treatments (described in greater detail below), these quantities are calculated by averaging over all of the other factor-level combinations, these quantities can be interpreted as conceptually similar to the main effects from a factorial experiment. As is the case in general with conjoint experiments, our interest here is less in rejecting null hypotheses of no effect, and more in comparing the relative magnitude of AMCEs: what factors do observers weigh most heavily when assessing resolve, and what factors do they largely ignore?

Capabilities and interests

We begin by looking at the effect of state-level characteristics on perceptions of resolve, starting with the results for the capabilities and interests at stake, shown at the top of Figure 2. Consistent with Press’s (2005) “current calculus” hypothesis — which we examine in greater detail below — we see that observers indeed place great weight on an actor’s military capabilities and level of interests at stake in the dispute. States with powerful militaries are perceived as 16.3% more likely to stand firm than states with less powerful militaries, while states whose stakes in the dispute are described as high are seen as 14.9% more likely to stand firm than states whose stakes in the dispute are described as low.

The substantively large effects of these two characteristics is of interest not just because of their prominence in theories in IR, but also because evolutionary theorists argue that selection pressures have hard-wired humans and other animals with cognitive and behavioral mechanisms geared towards assessing the strength and interest of others (Maynard Smith, 1974; Parker, 1974; Parker and Rubenstein, 1981). Although we
Figure 2 plots Average Marginal Component Effects (AMCEs) with 95% clustered bootstrapped confidence intervals. Positive values indicate that the attribute increases the perceived likelihood that an actor will be more likely to stand firm than its counterpart in the dispute, while negative values indicate that the attribute decreases the perceived likelihood of an actor standing firm. All estimates should be interpreted relative to the “base category,” indicated by the point without horizontal bars at x=0. Thus, for example, democracies are perceived as 4% less likely to stand firm than dictatorships.
present these cues directly for participants rather than having them assess capabilities and interests themselves, it is theoretically sensible that they would rely heavily on these indicators when assessing resolve.

**Regime type**

Turning to the effects of varying a state’s regime type, the results show that, in contrast, with theories of democratic superiority in crisis bargaining, respondents saw democratic states as 4.2% less likely to stand firm than dictatorships; states with mixed regime types in between democracies and dictatorships were seen as 1.4% less likely to stand firm than dictatorships, but the difference between the two is not statistically significant. These findings are of interest given debates in IR theory between “democratic triumphalists” and “defeatists” (Desch, 2008), the former touting the superiority of democracies both in terms of their likelihood of winning the wars they fight (Lake, 1992) and how they choose which wars to get into (Reiter and Stam, 2002), and the latter emphasizing how the mercurial and idealistic public threatens the viability of democratic foreign policy (Kennan, 1951; Lippmann, 1955). In his model of the informational effects of democratic institutions in crisis bargaining, Schultz (1999) finds that foreign rivals should be more likely to back down when facing democracies in a dispute, out of the belief that democratic institutions and a free press should make democracies less likely to bluff. Our results here fail to find support for that view, and are consistent instead with the apprehension voiced by Kennan (1951). At least as far as members of the U.S. public are concerned, democratic states are less likely to stand firm than non-democratic ones.

**Leader characteristics**

In general, the results above showed substantively strong results for the effects of country-level characteristics on assessments of resolve. An examination of leader characteristics offers mixed results. Against work suggesting that new leaders have a greater incentive to stand firm and develop a reputation for resolve in their early years in office, we find that new leaders are perceived as slightly (1.5%) less likely to stand firm compared to ones who have been in power for many years. Its relatively weak substantive effect, however, suggests that experience plays less of a role in assessing resolve than the other attributes discussed above. Similarly weak are the effects of gender: participants are no more likely to attribute resolve to male leaders than female ones. The null results of gender are noteworthy given gendered conceptions of leadership qualities in military contexts, and could be due to perceived selection effects, in which female politicians who are able to come to power are seen as no different than their male counterparts.\(^{15}\)

The one leader-level characteristic that participants do use as a heuristic to predict resolve is military

---

\(^{15}\)It is also possible that the null results of the gender treatment are due to the subtlety of the treatment, but since experimental work on prejudice and discrimination in other contexts employs manipulations of similar dosage, we find the selection effect argument more plausible.
experience: countries governed by leaders with extensive military experience are seen as 7.8% more resolved than those without any military experience; even those leaders with only brief time in the military get a boost in perceived resolve, of 3.3%. These results are thus consistent with work by Horowitz and Stam (2012) showing the distinctive foreign policy behavior of leaders with military experience, although unlike in their work, we do not distinguish between leaders with military backgrounds who have combat experience, and leaders with military backgrounds who do not. In general, though, the situational features of the crisis in terms of the balance of capabilities and interests between the belligerents contributes more to perceptions of resolve than characteristics of leaders themselves. Regardless of the extent to which leaders matter in international politics, it is of interest that at least with the characteristics we manipulate here, ordinary citizens tend to ascribe a greater role to broader structural factors than leader characteristics.

Past behavior

Past behavior represents a theoretically important set of attributes. In the experiment, we manipulated the country’s previous behavior in a foreign policy dispute in four different ways: (i) whether it initiated the crisis or was instead the target, (ii) whether the opponent the country faced in its previous crisis was itself an ally or adversary of the United States, (iii) whether the leader of the country at the time of the dispute is the same leader as the one in the current crisis, and (iv) whether in the previous crisis the country ultimately stood firm or backed down. In supplementary analyses in the appendix we interact all four of these past behavior variables with one another, but for ease of interpretation in the analysis presented in Figure 2 we reduce the number of quantities of interest and only interact the previous outcome and leader identity treatments, while presenting the average effects of the target/initiator and against ally/against adversary treatments. While neither of the latter significantly affects perceptions of resolve, we see theoretically important effects for the former.

First, states that backed down in their previous dispute are seen as significantly less resolved in the current crisis, offering evidence that participants are indeed drawing reputational inferences from past behavior. If the leader in charge was the same leader as in the current crisis, backing down corresponds with an 11% decrease in the perceived likelihood of standing firm, while a different leader backing down causes a 9.1% decrease in the perceived likelihood of standing firm. Thus, although it appears that backing down in past disputes is more informative under the same leader than a different one, the differences between the two effects are not statistically significant. In contrast, when the country stood firm in the previous dispute under the same leader, it is perceived as being 4.9% more likely to stand firm than when the country stood firm in the previous dispute under a different leader. Thus, we find evidence of leader-specific reputations for standing firm, but not for backing down. These results are sensible — consistent with an analogical
reasoning model of reputational inference in which observers draw stronger inferences from past behavior
the more the previous context resembles the current one (Shannon and Dennis, 2007) — but not trivial.16
Jervis (1982, 12-13), for example, notes a reputation paradox in which states that have backed down in the
past may be more likely to stand firm in the future precisely to avoid incurring reputation costs, which raises
the prospect that observers will expect states “to follow retreats with displays of firmness”, a pattern we do
not see here.

The results also challenge the Current Calculus hypothesis, which we operationalize by interacting a
state’s capabilities, interests, and outcome of the previous conflict. Although observers indeed look towards
capabilities and interests in calculating capability, past behavior still matters, and observers do draw infer-
ences from previous actions. As the quantities in Figure 3(a) illustrate, at every combination of capabilities
and interests, participants see standing firm in the past as significantly boosting the likelihood of standing
firm in the present, and the informative value of past behavior does not vary across different levels of current
capabilities and interests.17

Finally, Figure 3(b) tests the attribution theory hypothesis which holds that observers are more likely
to make situational attributions for desirable actions, and dispositional ones for undesirable ones, thereby
implying that adversaries will be unable to lose reputations for resolve by backing down, while allies will be
unable to gain reputations for resolve by standing firm. We operationalize this hypothesis by interacting the
outcome of the previous crisis with whether the country was an ally or adversary of the United States, and
holding the two relevant quantities of interest.

As the figure illustrates, we fail to find evidence in favor of the attribution theory hypothesis: allies who
stood firm in the past indeed gain reputations for resolve and are seen as more likely to stand firm in the
current crisis, while adversaries who backed down in the past indeed gain reputations for irresoluteness and
are seen as less likely to stand firm in the current crisis. In supplementary analyses, we operationalize the
concept of “desirability” further by interacting the ally/adversary and past outcome treatments with the
identity of the other state in the previous dispute, letting us explicitly test whether adversaries who back
down against allies in the past develop different reputations for resolve than adversaries who backed down
against other adversaries, and so on. The results hold, indicating that the effects of standing firm or backing
down in the past is not moderated by the relationship of the previous opponent in the dispute to the United
States.

16Following an analogical reasoning model, we would expect to find past behavior to be seen as even more predictive of future
behavior in cases where the opponent in the previous crisis is the same as in the current one.
17Indeed, BIC scores and likelihood ratio tests suggest the interactive model does not fit the data significantly better than
an additive one.
The figures depict Average Marginal Component Effects (AMCEs) with 95% clustered bootstrapped confidence intervals. Positive values indicate that the combination of attributes increases the perceived likelihood that an actor will stand firm, while negative values indicate that the combination of attributes decreases the perceived likelihood of an actor standing firm. The results in panel (a) offer limited support for the current calculus hypothesis: actors do draw inferences about resolve based on whether an actor stood firm or backed down in the past, but its effects can be overcome: if an actor backed down in the past dispute but its capabilities and stakes are both high, observers will still perceive it as being significantly more resolved than an actor that stood firm in the past dispute and has high capabilities but low stakes. The same pattern of results is obtained if we only include cases where the actor makes a public threat, or only include cases where the actor either makes a public threat or mobilizes troops. The results in panel (b) fail to offer support for the attribution theory hypothesis: allies who stand firm indeed gain reputations for resolve, while adversaries who back down gain reputations for irresolution. The quantities of interest here are calculated by interacting whether the actor is an ally or adversary of the US by whether the actor stood firm or backed down in the previous crisis. The same pattern of results is obtained if we further interact the attributes by whether the country’s previous crisis was against an ally or adversary of the US.
Current behavior

We next turn to the effect of current behavior on perceptions of resolve. Consistent with the literature on the informative value of public threats — which tie leaders’ hands if they don’t follow through (Fearon, 1997) — we see that countries that make public threats are perceived as 6.4% more likely to stand firm. Interestingly, though, participants see sunk costs in terms of troop mobilizations to be a more credible signal, with states that mobilize troops seen as 12% more likely to stand firm. Thus, we find that while talk might be cheap, it’s not free: our participants see it as significantly less costly than other signals that leader can send.¹⁸

Figure 4: Testing the democratic credibility hypothesis: democracies are not seen as making more credible threats than dictatorships

![Figure 4: Testing the democratic credibility hypothesis](chart)

Figure 4 plots Average Marginal Component Effects (AMCEs) with 95% clustered bootstrapped confidence intervals. Positive values indicate that the combination of attributes increases the perceived likelihood that an actor will stand firm, while negative values indicate that the attribute decreases the perceived likelihood of an actor standing firm. The results fail to offer support for the democratic credibility hypothesis: although public threats increase the perceived resolve of both democracies and dictatorships relative to when actors do not issue threats, democracies do not get a bigger credibility boost from public threats than dictatorships do. The quantities of interest here are calculated by interacting an actor’s regime type with whether an actor issued a public threat or not in the foreign policy crisis, dropping those cases where the actor mobilized troops instead (a sunk cost rather than hands-tying signal, about which democratic credibility theory is agnostic).

Finally, Figure 4 tests the democratic credibility hypothesis which holds that democracies — compared to their non-democratic counterparts — are better able to harness the signaling advantages of domestic audience costs and issue more credible threats. We operationalize the hypothesis by dropping those observations where

¹⁸Sinking costs and tying hands are ideal types of costly signals, but in reality many forms of signals involve a combination of the two. For example, a state that it is publicly mobilizing its troops can be seen not only as incurring sunk costs, but also as tying its hands, since the state can incur reputational costs for backing down. Similarly, as Slantchev (2005) argues, such military signals can also create bargaining advantages for the mobilizing side by shifting the balance of power in the crisis situation; our results are thus consistent with Sechser and Post (2015) on the superiority of muscle-flexing over hands tying.
countries mobilize troops — leaving only those where countries either did nothing or issued a threat — and interact the country’s regime type with whether it issued a public threat. Our key quantities of interest are bolded in Figure 4.

The results show that regimes of all types are able to credibly threaten. Importantly, though, democracies do not appear to display any unique advantages in this regard: democracies who issue threats are not seen as any more likely to stand firm than dictatorships who issue threats, nor is the effect of a public threat larger for democracies than for dictatorships. Our results are thus consistent with Weeks (2008), who argues that autocracies are also able to create audience costs, and Downes and Sechser (2012), who find that the democratic advantage in crisis bargaining is overstated.

In sum, then, although the experiment manipulated a large number of countervailing factors, our participants assessed resolve in clear and consistent ways. Although the treatment contrasts were deliberately high — the regime type treatment, for example, compares democracies with dictatorships, the alliance treatment compares allies with adversaries — the stark nature of the treatments cannot account for the results we observe: while all of our contrasts were designed to mirror significant real-world differences, some of them mattered a great deal while others did not. Moreover, differences between our subjects did not translate into differential responses to our treatments. In Appendix §1.3, we show that there is a conspicuous absence of heterogenous treatment effects: all types of respondents — more and less educated, hawks and doves, etc. — used the same heuristics to infer resolve. Moreover, we tested for and found scant evidence of higher-order interactions (Appendix §1.32); that is, the effect of any given variable, such as “capabilities” did not depend on the particular combination of other variables seen in a given choice task.

Interestingly, the four largest effect ranges in our analysis — capabilities, interests, past actions (particularly under the same leader) and costly signals — come from both the characteristic and behavioral cluster of indicators, suggesting that observers do not selectively attend to static over fluid indicators, or vice versa. More important, though, is that the four indicators observers saw as the most informative when assessing resolve are precisely the ones that deterrence scholars have traditionally highlighted (e.g., Schelling, 1960; Huth and Russett, 1984). Since we can safely presume our participants are not trained IR scholars, this raises interesting questions about why exactly ordinary members of the mass public carry around intuitive versions of deterrence theory in their heads — a point we return to below. Before we do, however, it is worth exploring whether elite decision-makers assess resolve in similar ways.
Generalizing to Leaders

While much is made of the question of “external validity,” and in particular the samples used, in experimental IR, it is important neither to overstate the problem nor to oversell the solution. Absent “strong prior beliefs that a theoretical-relevant difference exists” between students, the mass public and elite leaders, and that we know the direction of the resulting bias, there are serious tradeoffs to elite experiments (Renshon, Lee and Tingley, 2017; see also Renshon, 2017, 78-80). One obvious tradeoff is the complexity of such experiments; given sample size limitations, elite studies will never be as complex, comprehensive or subtle as their counterparts. However, in this case, there are two critical benefits to a study of resolve among elites: it both accomplishes an important methodological goal by conceptually replicating results from our conjoint experiment on the mass public, and extends the results to a population that is often (though not always) the focus of our IR theories.

To that end, we describe here the results of a second set of experiments (“Regime Type” and “Costly Signals”) conducted by the authors on an unusually elite sample of 89 current and former members of the Israeli Knesset (MKs). These leaders are among the most “elite” subjects yet recruited for IR studies, ranging all the way up to that of Prime Minister, with two-thirds of the sample having served on the foreign affairs and defense committee. While our conjoint experiment examined numerous factors, here we focus on two in particular: one from our characteristics (regime type) and the other from our behavior (costly signals) categories.

In the “Regime Type” experiment, subjects read about a situation in which two countries were involved in a public dispute over a contested territory. Country B was described as a dictatorship, while A was described as either a democracy or a dictatorship, controlling for many of the same factors manipulated in the conjoint experiment in the United States. We then asked, “Given the information available, what is your best estimate about whether Country A will stand firm in this dispute, ranging from 0% to 100%?” In this way, it was strikingly similar in its basic setup to the conjoint experiment: respondents were asked to play the part of observers to some conflict between two countries, given information about those two countries and then asked to estimate who would stand firm in the interstate dispute.

In our conjoint experiment on the U.S. public, our respondents saw democratic states as 4.2% less likely to stand firm than dictatorships. In the experiment conducted on Israeli Knesset members, as the bootstrapped distributions in the left-hand panel of Figure 5 shows, we obtained nearly identical results, with the elite sample perceiving democracies as around 6% less likely to stand firm than dictatorships.

---

19 Each experiment is itself the focus of a separate paper. Due to space constraints, we present here only “top-line” results from the parts of each study that relate to the conjoint experiment that is the focus of this paper, in order to provide an elite benchmark from which to evaluate the results. Further information can be found in Appendix §5.

20 The full text of both scenarios is contained in Appendix §5.
Figure 5: Comparing our mass public results with those from foreign decision-makers

Figure 5 compares the bootstrapped average treatment effects for three factors from the conjoint experiment, in blue (calculated using B=2500 clustered bootstraps), with the bootstrapped average treatment effects for the same three factors in a pair of survey experiments fielded on an elite sample of members of the Israeli Knesset, in red (calculated using B = 2500 bootstraps). We find strikingly similar findings across the studies, with mobilization being the only factor where the effect estimates significantly differ; although the elite distributions are fatter as a result of the much smaller sample size, supplementary analysis in Appendix §5 shows that when we downsample the public results to render the sample sizes more comparable between the two studies, the shape of the pairs of distributions becomes much more similar.

In the “Costly Signals” experiment, those same Knesset leaders were presented with a vignette that described a dispute between Israel and another country and asked to estimate the odds that the other country (Country B) would stand firm in the dispute. After that outcome question, which functioned as each subject’s baseline estimate of resolve, all subjects then read further text describing another version of the scenario in which Country B either made a public threat or mobilized their military. The study thus combined both within- and between-subjects designs. The former comes from each subject being in both a control (baseline) and treatment condition, while the latter comes from randomly assigning subjects to either the MOBILIZATION (sinking costs) or PUBLIC THREAT (tying hands) condition following the baseline scenario.

In our conjoint experiment on the U.S. public, we found that countries that make public threats are perceived as 6.4% more likely to stand firm, while troop mobilizations (sunk costs) increase estimates of resolve by 12%. As the middle and right-hand panel of Figure 5 shows, in our Knesset “Costly Signals” experiment, we once again found convergent results. Elite decision makers saw public threats as 8.1% more credible than the baseline estimates of resolve, while troop mobilization increased those estimates by 6.8%; of all three treatments in common between the three experiments, troop mobilization is the only one whose effect estimate significantly differs between the US mass public sample and the Israeli elite sample.

Half of the subjects read about Country B’s President issuing a public statement through the news media warning that they will “do whatever it takes to win this dispute” while the other half read about Country B “mobilizing their military and sending additional gunboats to the location of the dispute at sea.”
although the difference is one of magnitude rather than one of sign. Across both experiments, then, we
thus find evidence that public threats and sinking costs serve important and useful functions in effectively
signaling resolve to audience of leaders and publics alike. This is not to claim that elites and masses are
interchangeable, but merely to note that when we compare our experimental findings across the two samples,
they are more similar than different. Like other recent elite experimental work (Hafner-Burton et al., 2014;
Loewen et al., 2014; Renshon, 2015), then, our results caution against a reflexive “elite exceptionalism”
(Kertzer, 2016, 160-162), in which elites and masses are assumed ex ante to employ fundamentally different
cognitive architectures.

Conclusion

Much of the literature on reputation and resolve has been dominated by the question of what indicators we
use to assess resolve, and whether some indicators are more salient than others. Given a strategic environment
in which many indicators are available simultaneously and often lead to contradictory conclusions, observers
confront difficult decisions about which indicators to focus on and which to ignore. Indeed, the myriad
theories of resolve extant in IR — reputation, sinking costs, etc. — can be recast as explicit or implicit
suggestions as to which cues actors rely on when forming judgments about resolve.

Conceptually, extant studies on resolve inferences cluster around two families of indicators or heuristics:
behavioral indicators of resolve (in particular, the past and current crisis behavior of states) and several state-
level (i.e., capabilities, interests, alliances and regimes type) and leader-specific (i.e., leaders’ experience, time
in office and gender) characteristics. Using a conjoint experimental design we tested the relative importance of
particular behavior-based and characteristic-based indicators in shaping observers’ inferences about resolve
of adversaries and allies — and in so doing, evaluated the observable implications from a wide range of
theoretical frameworks across the discipline.

Our analysis allows us to distill several lessons, the first of which is that members of the public —
despite lacking specialized training, expertise or knowledge — are “intuitive deterrence theorists.” One
striking pattern in our results was the importance of capabilities, stakes, mobilization, and past actions, the
four variables with the largest substantive effects: a state whose stakes were high in a crisis was seen as
15% more likely to stand firm than a state with a lower level of interest. Notably, these four factors are
also the core ingredients of rational deterrence theory. Ordinary citizens, without ever having read Brodie
(1959) or Snyder (1961), seem to intuitively carry a “folk” version of deterrence theory around in their heads

22 We use “deterrence theory” here generally, such that the factors we discuss here are as relevant in cases of compellence as
whether because of socialization, or because assessing resolve resembles a basic adaptive problem in human evolution, for example — is obviously beyond the scope of this study, but two patterns are striking. First, in Appendix §1.3.1-1.3.2, we show that there is a conspicuous absence of heterogeneity in the experimental results: all types of respondents — more and less educated, hawks and doves, etc. — seem to rely on the same cues to the same extent; there is a similarly conspicuous absence of interactions between treatments, such that people seem to anchor on the same cues regardless of the particular combinations displayed; as Appendix §1 shows, there are also no signs of demand effects, in that participants’ responses are stable over time. Second, in Appendix §1.4, we examine changes in response times between disjunctive and conjunctive treatment assignments to demonstrate the extent to which the cues that display the largest treatment effects also seem to simplify participants’ decision process. Together, these patterns suggest the existence of a relatively ubiquitous mental model, regardless of its origin.

A second lesson concerns one of IR’s most important state-level attributes, regime type. Contrary to theories of democratic triumphalism, observers tend to see democracies as less resolved than dictatorships, rather than the other way around. This was the case in our mass sample of the U.S. public as well as in our novel sample of Israeli leaders. And while first-generation work on audience costs suggested that democracies would have a distinct advantage in this method of signaling, we do not find any evidence in support of this notion. While countries are able to use public threats to increase estimates of their likelihood of standing firm — and costlier actions such as mobilizing troops are correspondingly more effective — our results favor the theory of Weeks’ (2008) and her claim that democracies and autocracies do not differ in their abilities to generate audience costs.

A third lesson concerns the critical importance of past actions. In favor of the “confident theoretical beliefs” of many IR scholars (Dafoe, Renshon and Huth, 2014, 384) and against the “reputation paradox,” we find strong evidence that past actions affect current estimates of resolve; past actions speak about as loudly as present ones. And, in keeping with other work on this topic, we find additional evidence that reputations do not attach only to the state; in fact, past actions undertaken with the same leader are more informative of current resolve than past actions taken prior to a leadership turnover (Renshon, Dafoe and Huth, Forthcoming). By leveraging our conjoint experimental design, we were also able to shed light on more complex, multi-dimensional theories of resolve and reputation. For example, we found evidence inconsistent with Press’ (2005) “current calculus” theory of resolve: observers look to interests and capabilities as heuristics in assessing resolve, but past behavior matters as well.

Although our elite experimental evidence lacked the multidimensional richness provided by the conjoint

---

Disentangling these two mechanisms is particularly difficult given the extent to which classic works in deterrence theory, for example, rely on analogies to daily family life in order to explicate their claims - e.g., Schelling 1960.
design — though they control for many of the same factors that the conjoint experiment manipulates — the results they uncover are strikingly similar, not just in direction, but also in magnitude. These results cannot speak to the question of intuitive deterrence theory or the relative importance of different types of cues, but the congruence reported in the results are a useful corrective against the automatic assumption that leaders rely on a profoundly different cognitive architectures than ordinary citizens — a claim that merits empirical testing, rather than uncritical acceptance.

As in any experiment, some of the results we see here are likely shaped by the current international political climate: participants’ democratic defeatism, for example, reflects an era in which authoritarian states are increasingly flexing their muscles on the world stage (Gat, 2007). Yet it is similarly difficult to disentangle IR scholars’ enchantment with democratic triumphalism throughout the 1990s from the buoyant “end of history” in which they were writing. In this sense, we might expect that the weight observers place on some of the factors we explore here are likely to shift over time. Future work should also explore how they shift across space; the fact that we detect democratic defeatism among both American and Israeli respondents, for example, who are likely to be highly socialized to venerate the superiority of democratic political institutions, suggests we might expect even stronger negative effects for regime type in other states.

Finally, experimental designs such as this one can be particularly useful in formalizing, isolating and testing multi-dimensional theories in IR, without the concerns about endogeneity, collinearity, and aggregation that IR scholars must wrestle with when they use observational data (Tomz and Weeks, 2013). We drew from a varied and rich literature in reputations, credibility and resolve in IR to construct our conceptual framework. However, it would be a shame were the feedback loop to end here. One of the most useful directions for future research on this topic — in addition to conceptual replications and extensions of this experiment — would be for the results to feed back into future case studies on assessing resolve in IR. In that way, our theories of resolve will accumulate the advantages of each particular method and data source while minimizing the harms of each method’s weaknesses.

References


Bak, Daehee and Glenn Palmer. 2010. “Leader Tenure, Age, and International Conflict.” Foreign Policy


Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto. 2015. “Validating vignette and conjoint survey experiments against real-world behavior.” *Proceedings of the National Academy of Sciences*.


URL: http://sites.duke.edu/2014bmp/files/2014/10/Loewen_et_al.pdf


May, Ernest R. 1961. Imperial Diplomacy: The Emergence of America as a Great Power. New York:


Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton, NJ: Princeton University Press.
1 Robustness checks 3
1.1 Dropping unusual dyadic combinations ................................. 4
   Figure 1: Randomization check ........................................ 5
1.2 Fully saturated interactions for past behavior .......................... 6
   Figure 1: Robustness check: dropping the sixth choice task .......... 7
   Figure 2: Model diagnostics: testing the profile order assumption .... 8
   Figure 3: Model diagnostics: treatment effects by attribute order .... 9
   Figure 4: Robustness check: dropping unusual combinations .......... 10
   Figure 5: Robustness check: fully-saturated past action interactions 11
1.3 Estimating heterogeneous treatment effects ............................. 12
   1.3.1 By respondent characteristics ...................................... 12
   Figure 6: Diagnostics: heterogeneous treatment effects for education 13
   Figure 7: Diagnostics: heterogeneous treatment effects for hawkishness 14
   Figure 8: Diagnostics: heterogeneous treatment effects for partisanship 15
   Figure 9: Diagnostics: heterogeneous treatment effects for foreign policy interest 16
   1.3.2 Average marginal treatment interaction effects (AMTIEs) ......... 17
   Figure 10: Estimated Ranges of AMTIEs (I) .............................. 18
   Figure 11: Estimated Ranges of AMTIEs (II) ............................. 19
1.4 Response latency results .................................................... 20
   Figure 12: Response latency effects ....................................... 22

2 Dispositional Instrument 23
2.1 Militant Internationalism .................................................... 23
2.2 Cooperative internationalism .............................................. 23
2.3 Isolationalism .................................................................... 23
2.4 International trust ............................................................. 23
2.5 Demographic Questions ....................................................... 24

3 Conjoint Instrument Screen .................................................... 25

4 Sample information, reweighting, and demand effects 25
   Table 3: Sample characteristics .............................................. 26
   Figure 13: Comparing weighted and unweighted AMCEs ................ 28

5 Knesset Studies 29
5.1 Recruitment protocol .......................................................... 29
   5.1.1 Participant verification protocol ......................................... 30
   Figure 14: Recruitment Letter .................................................. 31
5.2 Sample characteristics and representativeness ............................ 32
   Table 4: Knesset Sample Characteristics .................................... 33
   Table 5: Sample representativeness tests ................................. 34
5.3 Supplementary results ......................................................... 34
Figure 15: Downsampling public results ..................................... 35
5.4 Experiment Instrumentation .................................................. 36
5.4.1 Regime Type Experiment .................................................. 36
5.4.2 Costly Signals Experiment ................................................ 36
1 Robustness checks

As we note in the main text, conjoint experimental designs rest on a small number of assumptions. First, the stability and carryover effects assumption, which mandates that potential outcomes remain stable across periods — that is, a participant should choose a given country (conditional on its and the other country’s attributes) regardless of what countries or choices they had seen previously. This is not only important to ensure that the AMCE is a meaningful quantity of interest, but also is important to guard against the potential for demand effects, in which participants might respond differently over multiple rounds as they become familiar with the purpose of the study. Second, the no profile order effects assumption, which posits that respondents’ choices would be the same regardless of the order in which the two countries are presented within a choice task. Third is successful randomization, which simply assumes that the attributes of each profile are randomly generated. We test each of these assumptions in turn.

First, re-estimating the quantities of interest within each round suggest that the AMCEs are stable across rounds, except for in the sixth round, where the treatment effects are noticeably smaller in that round than in all of its counterparts. These temporarily smaller effect sizes are difficult to explain theoretically and are unlikely to be due to respondent fatigue, since the treatment effects in subsequent rounds return to their previous levels. We replicate the results from the main analysis in Figure 1, this time dropping the sixth round, and find that the results remain the same. Since including the sixth round doesn’t substantively change our results (and if anything, renders them slightly more conservative), we include all eight rounds in all of the empirical results reported in the main text. This intertemporal stability also mitigates any potential concerns about demand effects, as our participants respond to the treatments in the last round much as they do in the first. We return to this point in Appendix §4.

Second, Figure 2 tests the profile order assumption, showing that the results do not appear to systematically differ based on whether a particular characteristic was presented as belonging to country A or country B. Third, Table 1 presents the results from our randomization check, showing that the treatment assignments are relatively well-balanced across a host of demographic characteristics. Finally, Figure 3 tests whether the row order in which attributes were presented to respondents within the experimental stimuli changes their results. Recall that although the order of attributes was randomized across respondents, it was held constant across rounds for any given
respondent: that is, if a participant saw regime type in the first row of the table in the first round, that participant saw regime type in the first row of the table throughout all seven subsequent rounds. Figure 3 shows that there do not appear to be systematic differences in the AMCEs by attribute order: characteristics presented in the first row are not significantly stronger than those in the last row, for example, and characteristics presented in the first and last rows are not systematically different from those presented in the middle. Interestingly, although the row order randomization is designed to prevent treatments featured more prominently in the table from having a stronger effect, in the "real world", media outlets, political entrepreneurs, and political elites are likely to play a role in ensuring people are more likely to receive some treatments than others. Future work could therefore benefit from exploring these added layers to the information environment in which observers are situated.

1.1 Dropping unusual dyadic combinations

As noted in the main text, the analyses presented above randomly generated characteristics of both countries in each dispute, such that the leader characteristics of one country, for example, are independent of the leader characteristics of the other. To prevent odd combinations of actor characteristics biasing the experimental results, we impose some randomization constraints: if one of the countries in the dispute is randomly assigned to be the United States, randomization was constrained for other aspects of the country, as well as its opponent: if a country was assigned to be the United States, for example, it was described as being a democracy and having a very powerful military, while the identity of the opponent was constrained such that it could not also be the United States.¹

Furthermore, to preclude the possibility that unusual dyadic combinations of characteristics are skewing the results, Figure 4 below replicates the results from the main set of analyses, but dropping all disputes where two democracies faced off against one another, or two allies of the United States faced off against one another. A comparison of this restricted set of observations (in grey) with the unrestricted set from the main analyses (in black) shows that the results hold regardless of whether the dyads are included or not — it appears that allies are seen as relatively more resolute in the restricted model than the unrestricted one, but the difference is not statistically significant.

¹Ultimately, however, as noted in the main text, we decided to drop these observations from the study, to maintain the paper’s focus on how observers assess resolve, rather than also introducing the question of how participants do so.
Table 1: Randomization check

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High capabilities</td>
<td>-0.173, 0.129</td>
<td>-0.173, 0.128</td>
<td>-0.05, 0.269</td>
<td>-0.178, 0.114</td>
</tr>
<tr>
<td>Male</td>
<td>-0.095, 0.032</td>
<td>-0.028, 0.1</td>
<td>-0.055, 0.076</td>
<td>-0.014, 0.111</td>
</tr>
<tr>
<td>Age</td>
<td>-0.004, 0.002</td>
<td>-0.001, 0.004</td>
<td>-0.003, 0.002</td>
<td>-0.003, 0.002</td>
</tr>
<tr>
<td>Education</td>
<td>-0.021, 0.024</td>
<td>-0.042, 0.001</td>
<td>-0.032, 0.012</td>
<td>-0.016, 0.029</td>
</tr>
<tr>
<td>Party ID</td>
<td>-0.021, 0.233</td>
<td>-0.137, 0.104</td>
<td>-0.133, 0.128</td>
<td>-0.151, 0.096</td>
</tr>
<tr>
<td>Mil Assert</td>
<td>-0.148, 0.192</td>
<td>-0.091, 0.223</td>
<td>-0.278, 0.031</td>
<td>-0.172, 0.152</td>
</tr>
</tbody>
</table>

|                  | (5)                  | (6)                  | (7)                  | (8)                  |
| Adversary        | -0.002, 0.28         | -0.182, 0.118        | -0.115, 0.201        | -0.15, 0.142         |
| Against adversary| -0.039, 0.089        | -0.046, 0.078        | -0.084, 0.037        | -0.034, 0.091        |
| Initiator        | -0.003, 0.003        | -0.002, 0.003        | -0.004, 0.002        | -0.006, 0            |
| Backed down      | -0.043, 0.003        | -0.021, 0.026        | -0.024, 0.022        | -0.011, 0.034        |
| Party ID         | -0.265, -0.001       | -0.149, 0.104        | -0.232, 0.027        | -0.082, 0.18         |
| Mil Assert       | -0.206, 0.115        | -0.162, 0.153        | -0.025, 0.32         | -0.075, 0.262        |

|                  | (9)                  | (10)                 | (11)                 | (12)                 |
| Same leader      | -0.09, 0.21          | -0.291, 0.087        | -0.341, 0.001        | -0.087, 0.269        |
| Democracy        | -0.052, 0.07         | -0.018, 0.131        | 0.027, 0.172         | -0.121, 0.028        |
| Mixed            | -0.006, 0            | -0.001, 0.006        | 0, 0.007            | -0.006, 0.002        |
| High service     | -0.02, 0.024         | -0.022, 0.034        | -0.021, 0.038        | -0.031, 0.025        |
| Party ID         | -0.058, 0.193        | -0.191, 0.116        | -0.141, 0.139        | 0.001, 0.292         |
| Mil Assert       | -0.168, 0.164        | -0.222, 0.134        | -0.235, 0.147        | -0.354, 0.051        |

|                  | (13)                 | (14)                 | (15)                 |
| Some service     | -0.107, 0.256        | -0.221, 0.132        | -0.167, 0.197        |
| Mobilized troops | -0.06, 0.091         | -0.015, 0.144        | -0.019, 0.135        |
| Public threat    | -0.001, 0.005        | -0.002, 0.004        | -0.001, 0.005        |
| Party ID         | -0.051, 0.003        | -0.053, -0.001       | -0.069, -0.014       |
| Mil Assert       | -0.423, -0.025       | -0.093, 0.306        | -0.113, 0.282        |

Note: Models 1-9 depict quantile-based clustered bootstrapped 95% confidence intervals from a series of logistic regression models, while models 10-15 depict the quantile-based clustered bootstrapped 95% confidence intervals from a series of multinomial logit models. The results show the treatment assignment is relatively well-balanced.
1.2 Fully saturated interactions for past behavior

The analysis of the effects of past behavior in the main text presents interactions between past actions (stood firm versus backed down) and the identity of the leader responsible (a different leader than in the current dispute, versus the same leader as in the current dispute), but presents the other two past action treatments (whether the state was the target or initiator, and whether the opponent in the previous crisis was an ally or adversary of the US) as average effects rather than estimate a more complex four-way interaction. Figure 5 replicates the results from the main text, but this time with the fully-saturated four-way interaction. The results are harder to interpret because of the larger number of moving parts, but lends itself to the same substantive conclusion: behavior in a previous crisis is seen as more informative of present behavior if led by the same leader as in the current crisis than a different leader. In contrast, whether the state was the target or initiator, or was facing an ally or adversary of the US, does not appear to meaningfully interact with past actions in shaping assessments of resolve.
To test the stability and no carryover effects assumption, Figure 1 plots Average Marginal Component Effects (AMCEs) with 95% clustered bootstrapped confidence intervals, replicating Figure ?? from the main text, but without including data from the sixth choice task, which displayed violations of the carryover assumption. Importantly, the results hold, and are stronger than the more conservative estimates presented in the text. As before, positive values indicate that the attribute increases the perceived likelihood that an actor will stand firm, while negative values indicate that the attribute decreases the perceived likelihood of an actor standing firm. Thus, for example, democracies are perceived as 4% less likely to stand firm than dictatorships.
Figure 2 plots Average Marginal Component Effects (AMCEs) with 95% clustered bootstrapped confidence intervals, replicating the results from the main text, but conditional on whether each characteristic was presented as belonging to country A (in black) or country B (in grey). Importantly, the results do not appear to systematically differ based on whether a particular characteristic was presented first or second.
Figure 3: Model diagnostics: treatment effects do not systematically vary by attribute order
Figure 4: Dropping unusual dyadic combinations
Figure 5: Estimating the full four-way interaction for past actions
1.3 Estimating heterogeneous treatment effects

1.3.1 By respondent characteristics

Figures 6-9 look for heterogeneous treatment effects across four different sets of characteristics: education (do respondents with a college degree rely on different cues than those without one?), militant internationalism (do hawks use different cues than doves?), partisanship (do Republicans use different cues than Democrats?), and interest in foreign affairs (do particularly engaged respondents employ different heuristics than respondents who are not as interested in international politics?).

The results find relatively little evidence of treatment heterogeneity: in Figure 6 the low-educated respondents (in grey) and high-educated respondents (in black) seem to use remarkably consistent heuristics: more educated respondents are slightly more pessimistic about democracies, for example, but the overall results remain the same. Similarly, in Figure 7, doves (in grey) and hawks (in black) also display strikingly similar results. Hawks appear less likely to give dictatorships the benefit of the doubt, and see allies as more reliable, but these differences are noticeably small given the vast differences between doves and haws we find in other areas of public opinion about foreign policy.

We might expect stronger treatment heterogeneity with respect to militant internationalism if the United States was itself a participant in the disputes, rather than merely an observer. Finally, although the confidence intervals around the point estimates in Figure 9 are wider for respondents with high levels of foreign policy interest (in black) than low levels of foreign policy interest (in grey), we see fairly similar results between the two: it appears that more interested respondents place a greater weight on military capabilities, and somewhat lesser weights on costly signals, but the differences for the latter are relatively small.

In general, then, there do not appear to be systematic differences across any of these characteristics: we never see that certain kinds of respondents are systematically more sensitive to leader-level variables rather than country-level ones, for example, or draw more information from current behavior and less from past behavior. The absence of these systematic differences is theoretically interesting, in that it shows a relatively uniform pattern of cue use across types of respondents, consistent with the notion of an “intuitive deterrence theory” articulated in the main text, in which even those who think about world politics in fundamentally different ways nonetheless seem to place similar weights on the same set of indicators.

\(^2\)We calculate the threshold for doves and hawks, and low and high interest in foreign policy by mean-splitting responses in the militant internationalism and foreign policy interest scales, shown in full below.
Figure 6: Heterogeneous treatment effects: low-educated (grey) versus high-educated (black)
Figure 7: Heterogeneous treatment effects: hawks (black) versus doves (grey)
Figure 8: Heterogeneous treatment effects: Democrats (grey) versus Republicans (black)
Figure 9: Heterogeneous treatment effects: low foreign policy interest (grey) versus high foreign policy interest (black)
1.3.2 Average marginal treatment interaction effects (AMTIEs)

The analysis above estimates heterogeneous treatment effects with respect to respondent characteristics, but we can also look for interaction effects between the treatments themselves. Because interpreting cross-treatment interaction effects in conjoint experiments is sensitive to the choice of the baseline category, we follow Egami and Imai (2015) in presenting Average Marginal Treatment Interaction Effects (AMTIEs). Figures 10-11 thus present the full-range of one-way, two-way, and three-way AMTIEs for each of the treatment categories presented in the paper, letting us test for the possibility of higher-order interactive effects between sets of treatments. Importantly, just as we find relatively little evidence of heterogeneous treatment effects with respect to respondent characteristics, we also find relatively little causal heterogeneity between treatment combinations. Figure 10(a) reconfirms the findings from the main analysis, showing that the largest treatment effects are capabilities, stakes, past actions (operationalized here, as in the main text, based on whether the country backed down or stood firm in the previous crisis, conditioned on whether a different leader was in power at the time), and current behavior (costly signaling). In contrast, Figure 10(b) shows that the magnitude of the two-way effects are much smaller, suggesting we lose little by focusing solely on the ACMEs in the main text. Indeed, 30 of the 55 two-way AMTIEs have effect estimates of 0, and are thus omitted from the plot to save space. In Figure 11(a), the three-way AMTIE estimates are presented; not only are they extremely small, but only five estimates are presented, because the other 160 three-way AMTIEs have effect estimates of 0. Figure 11(b) illustrates the same pattern a different way, presenting the largest five one-way, two-way and three-way AMTIEs, showing once again that we can safely ignore higher-level interactions. In this sense, the weak higher-order interactions also reconfirm the lack of evidence in these experimental results in favor of the the current calculus and attribution theory hypotheses, both of which posit interactions between past actions and other variables. Finally, these findings also extend those from the previous subsection, in that individuals seem to anchor on the same cues regardless of the specific combination of treatments presented: it is not that particular higher-order interactions make capabilities and interests matter any less, for example. This offers further evidence in support of our “intuitive deterrence theory” model.
Figure 10: Estimated Ranges of AMTIEs (I)

(a) One-way AMTIEs

(b) Two-way AMTIEs

Note: Panel (a) reconfirms the importance of capabilities, stakes, past actions, and current behavior (costly signaling), while other variables exert relatively weak effects. Panel (b) shows little evidence of significant two-way AMTIEs. 30 of the 55 two-way AMTIEs have effect estimates of 0, and are omitted here to save space.
Figure 11: Estimated Ranges of AMTIEs (II)

(c) Three-way AMTIEs

(d) Largest AMTIEs

Note: Panel (c) shows little evidence of significant three-way AMTIEs. 160 of the 165 three-way AMTIEs have effect estimates of 0, and are omitted here to save space. Panel (d) shows the five largest AMTIEs of each type.
1.4 Response latency results

The results presented above suggest the presence of a fairly ubiquitous mental model. Appendix §1.3.1 showed that when it comes to assessing resolve, participants with more education rely on a similar portfolio of cues as those with less education, participants more interested in foreign affairs assess resolve using a similar set of indicators as participants less interested in foreign affairs, and hawks utilize a similar set of cues as doves; Appendix §1.3.2 showed a similar absence of interactions between treatments, such that the effects of the cues with the largest effects in our results (capabilities and stakes, for instance) do not seem to systematically vary based on the presence of particular combinations of treatments. This absence of contingence is striking, in that it paints a picture of respondents as sharing a common schema (which in the main text we call “intuitive deterrence theory”) in which participants focus their attention on a particular set of cues (particularly capabilities and stakes) to resolve the ill-structured problem of assessing resolve in disputes. Although the strong effects of these AMCEs in the experimental results document this pattern nicely, another way of further confirming it involves the use of response times.

The logic of response latency analysis (Mulligan et al., 2003) is straightforward: the question of how observers assess resolve is a ultimately a question of how individuals process information, and one way scholars of political behavior indirectly get at information processing involves looking at the speed at which individuals take to answer a question after it has been asked (e.g. Bolsen, Druckman and Cook, 2014). Although response latencies are inherently noisy measures, if on average respondents presented with certain combinations of treatments take systematically longer or shorter to respond than respondents presented with other combinations of treatments, it potentially sheds light on the cognitive mechanisms under investigation, whether processing efficiency or attitude accessibility (Fazio, 1990). For our purpose, two tests are potentially instructive. One would involve comparing how average response latency changes as the number of factors manipulated changes. For example, if participants assess resolve more quickly when capabilities information is presented than when it isn’t (even though the amount of text participants are being presented with would actually increase!), this would offer suggestive evidence that participants rely on capabilities cues when assessing resolve. In the case of our experimental design, because none of our factors being manipulated has a pure control, we are unable to use such an estimation strategy here.\(^3\) Instead, we use a different test, exploiting the choice-based nature of our conjoint design to test for the effects of

\(^3\)For a broader discussion of this type of approach, see Acharya, Blackwell and Sen (2017).
conjunctive versus disjunctive treatment assignments between the two randomly generated profiles.

For purposes of simplicity, suppose four choice tasks, each between two country profiles, illustrated in Table 2.

### Table 2: Hypothetical treatment assignments

<table>
<thead>
<tr>
<th>Choice</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Factor 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Factor 3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Factor 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

For each choice task, each choice profile consists of four randomly generated factors, each with multiple levels. For Task 1, the treatment assignments between each profile are disjunctive, in the sense that each of the four factors are assigned to different levels between country A and country B. The same is true for Task 3. For Task 2, however, the treatment assignment for factor 2 is conjunctive, in that the same level of the treatment is assigned to both choices in the set. The same is true for Task 4. By comparing how the average response latency differs between the conjunctive and disjunctive choice sets for each treatment (e.g., for factor 2, a comparison of the average response latency for tasks 1 and 3, versus tasks 2 and 4), we thus have a sense of how much respondents anchor on a particular factor as part of their decision-making process. For example, in a model of consumer choice, if individuals take much longer to decide between a pair of consumer goods if both options have the same price, this implies that price is an important consideration. The same holds here: if capabilities or stakes are important considerations, participants should assess resolve more quickly when given conjunctive capability or stake treatments than when given conjunctive ones.

In Figure 12, we present point estimates and 95% clustered bootstrapped confidence intervals from a regression model where each participant’s logged response time is modeled as a function of conjunctive treatment assignments for each of the treatments from the conjoint design; each point estimate thus depicts a different coefficient estimate from the model. The plot shows that conjunctive cues for capabilities and stakes significantly increases response time, suggesting that a balance of power or balance of interests induces respondents to take longer when assessing resolve. Additional analysis confirms this interpretation: the average response time when both countries have low capabilities is indistinguishable from when they both have high capabilities, but the average response time when one country has high capabilities and the other country has low is approximately two seconds faster, showing how an imbalance of power accelerates the assessment process. A similar pattern emerges with respect to stakes: the average response time when both countries have low stakes is similar to when they both have high stakes, but the average response time when one country has high stakes and the other country has low stakes is roughly 1.5 seconds faster.
ingly, the other factor that displays conjunctive cue effects is the identity of the actor (an adversary or ally of the United States); this finding is striking given that the AMCE for the actor treatment is significant but relatively modest in size. The significant results with respect to logged response time are perhaps due to the centrality and automaticity of social categorization (Kurzban, Tooby and Cosmides, 2001) and importance of coalitional psychology (Lopez, McDermott and Petersen, 2011): even if people ultimately don’t consider whether an actor is an ally or adversary to be a particularly strong indicator of how much resolve an actor will display in a crisis, participants take roughly 1.5 seconds faster to assess resolve in disputes between a member of the ingroup and a member of the outgroup than they do in disputes between outgroup members or between ingroup members.
2 Dispositional Instrument

As noted in the main text, in addition to the choice-based conjoint experiments, participants also completed a battery of dispositional and demographic measures. To avoid downstream effects, participants were randomly assigned to receive the dispositional questionnaire either before completing the conjoint tasks, or afterwards. The instrumentation used below is relatively standard in the public opinion about foreign policy literature, based off of classic work by Holsti and Rosenau (1988) and Wittkopf (1990); see, e.g. Kertzer et al. (2014).

Unless otherwise specified, all response options below are scaled from “strongly agree” (1) to “strongly disagree” (5).

2.1 Militant Internationalism

1. The best way to ensure world peace is through American military strength

2. The use of force generally makes problems worse [reverse-coded]

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

2.2 Cooperative Internationalism

1. America needs to cooperate more with the United Nations in settling international disputes

2. It is essential for the United State to work with other nations to solve problems such as overpopulation, hunger and pollution

2.3 Isolationalism

1. The U.S. needs to play an active role in solving conflicts around the world [reverse-coded]

2. The U.S. government should just try to take care of the well-being of Americans and not get involved with other nations
2.4 International trust

1. Generally speaking, would you say that the United States can trust other nations, or that the United States can’t be too careful in dealing with other nations? [the United States can trust other nations/ the United States can’t be too careful in dealing with other nations]

2.5 Demographic Questions

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

2. What year were you born? [open text box]

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/ Master’s degree/ Doctoral degree/ Professional degree (e.g., JD or MD)]

4. Generally speaking, do you usually think of yourself as a Republican, Democrat, or as an independent (check the option that best applies)? [Strong Republican/ Republican/ Independent, but lean Republican/ Independent/ Independent, but lean Democrat/ Democrat/ Strong Democrat]

5. Below is a scale on which the political views that people might hold are arranged from “extremely conservative” to “extremely liberal.” Where would you place yourself on this scale? [extremely conservative/ conservative/ slightly conservative/ moderate/ slightly liberal/ liberal/ extremely liberal]

6. How interested are you in information about what’s going on in foreign affairs? [extremely interested/ very interested/ moderately interested/ slightly interested/ not interested at all]

7. How interested are you in information about what’s going on in government and politics? [extremely interested/ very interested/ moderately interested/ slightly interested/ not interested at all]
3 Conjoint Instrument Screen

In this portion of the study, we will present you with information about a series of eight foreign policy disputes between different countries.

Countries often get into disputes over contested territories. These disputes receive considerable attention because of the risk they can escalate to the use of force. Thus, the kinds of disputes described here are ones that have occurred many times, and will likely occur again.

In each screen, we will present you with a pair of countries involved in a territorial dispute, tell you a bit about each of them, and ask you to make predictions about what you think will happen. There are no right or wrong answers, we’re simply interested in the kinds of predictions you make.

4 Sample information, reweighting, and demand effects

As noted in the main paper, the study was fielded on a sample of 2009 American adults recruited via Amazon Mechanical Turk in January 2015. Participants were paid $1 for their participation. One potential concern about the use of MTurk is the representativeness of the sample, as compared to other potential survey platforms. However, Berinsky, Huber and Lenz (2012, 366) show that MTurk samples are “often more representative of the general population and substantially less expensive to recruit” than other “convenience samples” often used in political science (see Huff and Tingley, 2015, for the latest and most definitive work on this subject). 5 They also demonstrate the ability to replicate results from nationally-representative samples — e.g., Kam and Simon’s (2010) work on framing and risk and Tversky and Kahneman’s (1981) classic “Asian Disease problem” — using MTurk workers. 6 As a result, MTurk is becoming increasingly widely used in experimental political science, and experiments using MTurk samples have been published in a variety of notable journals, including the American Political Science Review (Tomz and Weeks, 2013), the American Journal of Political Science (Healy and Lenz, 2014; Levy et al., 2015; Bishin et al., 2016), Inter-

5 Though compared to nationally representative samples, MTurk workers tend to be younger and more ideologically liberal.

6 While there has been some initial wariness regarding online experiments, many famous and well-known behavioral studies have been replicated using MTurk. For more on this, see Mason and Suri (2010); Buhrmester, Kwang and Gosling (2011); Rand (2012). For a different viewpoint, see Krupnikov and Levine (2014), though their caution applies only to very specific kinds of experimental studies, as we discuss below.
national Organization (Wallace, 2013), and the Journal of Conflict Resolution (Kriner and Shen, 2014). In keeping with “best practices” suggested by numerous researchers, we limited participation in the study to MTurk workers located in the United States, who had completed ≥ 50 HITs, and whose HIT approval rate was >95%.

Table 3: Sample characteristics

<table>
<thead>
<tr>
<th>Sample Characteristic</th>
<th>Unweighted Sample</th>
<th>Weighted Sample</th>
<th>Population Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.541</td>
<td>0.487</td>
<td>0.492</td>
</tr>
<tr>
<td>18 to 24 years</td>
<td>0.136</td>
<td>0.138</td>
<td>0.128</td>
</tr>
<tr>
<td>25 to 44 years</td>
<td>0.676</td>
<td>0.361</td>
<td>0.342</td>
</tr>
<tr>
<td>45 to 64 years</td>
<td>0.167</td>
<td>0.311</td>
<td>0.341</td>
</tr>
<tr>
<td>65 years and over</td>
<td>0.021</td>
<td>0.189</td>
<td>0.189</td>
</tr>
<tr>
<td>High School or less</td>
<td>0.111</td>
<td>0.397</td>
<td>0.420</td>
</tr>
<tr>
<td>Some college</td>
<td>0.284</td>
<td>0.200</td>
<td>0.194</td>
</tr>
<tr>
<td>College/University</td>
<td>0.491</td>
<td>0.299</td>
<td>0.282</td>
</tr>
<tr>
<td>Graduate/Professional school</td>
<td>0.113</td>
<td>0.104</td>
<td>0.104</td>
</tr>
</tbody>
</table>

As Huff and Kertzer (Forthcoming) note, there are typically two concerns about the use of MTurk. The first involves the composition of the sample. As noted above, MTurk samples, although more diverse than most convenience samples used in political science, are nonetheless not nationally representative of the US population as a whole, tending to skew somewhat younger and more educated (Huff and Tingley, 2015). To address this point, we employ a two-pronged strategy. First, like Huff and Kertzer (Forthcoming), we use entropy balancing (Hainmueller, 2012) to reweight our sample towards national population parameters, trimming the weights to reduce the impact of extreme values. Table 3 presents the weighted and unweighted sample characteristics, showing that the weighted data hews closely towards population targets. Figure 13 replicates the results from the main text, but this time also superimposing the AMCEs from the weighted data. As before, capabilities, stakes, costly signals, and past actions are the most important indicators. The confidence intervals for regime type are wider, such that its effect is no longer significant, but a pairwise comparison indicates that the effects themselves do not significantly differ between the two datasets. Second, we present a series of models testing for heterogeneous treatment effects (Figures 6-9), which include a number of additional characteristics where we might expect our sample to differ but which are not explicitly being accounted for in the reweighting. For example, if participants recruited on MTurk also happen to be systematically more liberal, or more interested in politics, than the population at large, it is helpful to test whether more conservative participants, or less po-
politically engaged respondents, rely on systematically different heuristics. As the results in Appendix §1.3.1 show, we fail to find evidence that this is the case.

A second concern about MTurk might involve the potential for demand effects (Chandler, Mueller and Paolacci, 2014), the tendency for experimental participants to decipher the purpose of the study, and act in a way either consistent with the experimenters’ wishes (e.g. *experimenteer bias* - Orne, 1962; Rosenthal and Fode, 1963) or contrary to them (e.g. the *screw-you effect* - Masling, 1966). As is the case with participants in other online survey platforms, MTurk users often participate in a large number of studies, which not only raises concerns that study participants might be more susceptible to demand effects, but raises particular concerns for studies that require naive participants that have not encountered a particular experimental paradigm before (Paolacci and Chandler, 2014; Krupnikov and Levine, 2014; Huff and Kertzer, Forthcoming). Given the purpose of our study, these concerns are mitigated here. First, in an information-rich conjoint experiment with countervailing indicators, demand effects are less relevant than they might be in a less elaborate experimental design; it is presumably easier to determine which factor the experimenters are interested in in an experiment that manipulates one factor than an experiment that manipulates seventeen of them. Second, the within-subject component of conjoint designs minimizes the need for naive participants, since participants are being exposed to multiple treatment conditions across multiple rounds. Third, as noted in Appendix §1, we can test for demand effects explicitly with conjoint designs by validating the stability and carryover effect assumption. If demand effects are present, we should expect this assumption to be violated, as participants decipher the goals of the experimenters by taking multiple rounds of the experiments, and thus respond to the treatments differently over time. Instead, the results reported in Appendix §1 suggest the AMCEs are relatively stable over time, such that participants do not appear to respond any differently to the treatments in the last round as they do in the first. Thus, we have little reason to be concerned about demand effects here.
Figure 13: Comparing weighted and unweighted AMCEs

Figure 13 presents the unweighted point estimates and clustered bootstrapped confidence intervals from the main text (depicted by black solid circles) with their weighted counterparts (denoted by the open squares). Pairwise comparisons show the weighted results never significantly differ from the unweighed ones, although the confidence intervals for the weighted results are wider. As before, capabilities, stakes, costly signals, and past actions are the most important indicators.
5 Knesset Studies

Because of the 35 page limit in the main paper, below we present more supplementary information about the Knesset studies. We begin by discussing the recruitment procedure, before presenting the sample characteristics and representativeness tests; we then present some additional results, and the experiment instrumentation for both the regime type experiment, and the costly signals experiment.

5.1 Recruitment protocol

The recruitment process began by compiling a dataset of all 415 individuals who had served as members of Israeli Parliament (i.e., the Knesset) from the beginning of the 14th Knesset in June 1996 through the 20th Knesset (the current Knesset) that was sworn in in March 2015. We compiled a data set that included the following information about our population:

1. full name
2. party affiliation while in Knesset
3. names of all Knesset committees on which (s)he served
4. number of terms served
5. whether (s)he served as a minister in the government, and if so, what portfolios (s)he held
6. whether (s)he was a member of the Cabinet

Contact information for our participants was obtained through a variety of channels, including the Secretary of the Knesset, the Knesset Channel, the different parties’ leadership offices in the Knesset and other government offices where former Knesset members are currently employed. Email addresses for all current members of the Knesset were obtained through the Secretary of the Knesset. To verify whether the contact information we obtained was correct, we either called or emailed all the former Knesset members from the last twenty years and asked them if they would be interested in taking a “10 minute electronic survey by a team of professors from leading American Universities.” 30.6% of the initial population was removed from the sampling frame at this stage, either because the members were deceased, were too sick to participate, or because their contact information was out of date and newer contact information could not be found. This process left us with a sample
of 288 potential candidates to take our survey. This pool included all 120 current members of the
Knesset along with 168 former members whose contact information was available.

On July 10, 2015, we executed a soft launch of our on-line survey. The survey included a
recruitment email, written in Hebrew (reproduced below), a link to our on-line survey, and an
individual six-digit password that was pre-assigned to each member. In the following days, we
emailed the invitation to all current and former members in our dataset. A few weeks later, we sent
a reminder email to those who had not responded to the survey. We sent a third round of reminders
a few weeks later. In between these rounds, we phoned former and current Knesset members or their
assistants to remind them to take the survey. In early August, the Director of Academic Affairs at
the Knesset, together with the Secretary of the Knesset, sent an email to all current Knesset members
encouraging them to take the survey, repeating essentially the same information we provided in the
introductory email.

In addition to the on-line survey, we created identical hard-copy versions of our survey. In
mid-August we sent those who had not responded to our survey a reminder email and attached an
electronic copy of our survey that could be opened in Microsoft Word. Respondents were given the
option of either faxing or emailing the completed survey back to us. That same six-digit code was
the only identifying information on the paper copies of the survey, allowing us to track completion
among our sample population. Members of our research team also traveled to the Knesset on four
separate occasions to invite current members to participate.

The entire recruitment process was done in Hebrew. Two Hebrew-speaking research assistants
and one member of the research team who is a native Hebrew speaker corresponded with the members
of the Knesset or their assistants. Participants were informed that there would receive no financial
reward for taking the survey, but that we would be happy to share with them the results of the
survey. Moreover, participants were promised full anonymity: with the exception of the research
team, participants were assured that identifiable information would not be released or reported.

5.1.1 Participant verification protocol

We took several steps to increase our confidence that the current and former decision-makers partic-
ipated in the study rather than members of their staff. First, in the introductory email we explicitly
indicated that the questionnaire should be fielded by the decision-maker himself, and not by mem-
bers of his or her staff. We explained that the code we provided to access the on-line survey was
Figure 14: Recruitment Letter
personal, and should not be shared with others. Importantly, we did not offer any material incentives for filling out the survey, to dissuade decision-makers and assistants for taking the survey for those material reasons.

Second, in the survey itself we asked the participants to enter their complete date of birth. This allowed us to compare this information with the date of the decision-makers in official Knesset records. Third, for the 75% of our sample consisting of former Knesset members, a Hebrew-speaking research assistant and one of the authors were both in touch with the decision-maker directly via phone or email, and confirmed with him/her that they were the ones taking the survey. Anecdotally, our research team found that many of our participants were quite eager for the opportunity to opine on issues of foreign policy to an outside audience.

In the case of some current Knesset members, after receiving approval from their parliamentary assistant, a Hebrew-speaking research assistant from our team or one of the authors gave the Knesset members the survey directly and picked it up from them within a two-hour window. However, some Knesset members wished to maintain their anonymity and thus were not in direct contacted with the research team.

Finally, although we follow best practices, as is always the case with elite experiments, we should note that decision-makers who wished to “cheat” and delegate their participation to others could have probably found ways to do so. However, the combination of the types of questions asked in the survey, the absence of material compensation for survey completion, our explicit request the survey not be filled out by others, and the enthusiastic response to our survey from most of the decision-makers who took the survey leave us confident that the vast majority of them participated directly.

5.2 Sample characteristics and representativeness

Table 4 presents basic descriptive statistics for our Knesset sample. As the table shows, the sample is unusually “elite” by the standards of many experiments conducted in international relations: two-thirds of the sample has experience on the foreign affairs and defense committee, and over 40% served as deputy minister or higher. Moreover, because we have a defined population of elites from which we are sampling (see the recruitment protocol discussed above), we can also formally test how representative our participants are, along two dimensions. First, how do our participants compare to the complete population of individuals who served in the Knesset from 1996 to the
present? Second, how do our participants compare to our sampling frame, a different group than the complete population because it does not include members who had passed away, were too sick to participate, or for whom we were unable to acquire up to date contact information. Thus, whereas the first quantity explores whether our participants look like the universe of Knesset members in this time period, the second explores survey non-response. We explore both questions in Table 5 below, which presents a set of linear probability models comparing our participants to the universe of Knesset members from 1996-2015 (models 1-2) and to only those Knesset members who had been sent the survey (models 3-4). The results show that unsurprisingly, current members of the Knesset were less likely to participate in the survey than former members, but that interestingly, our participants are not significantly less ”elite”, as measured by the proportion of respondents with experience as deputy ministers, or as cabinet members or higher. If anything, our sample is slightly more experienced than the universe of decision-makers, though the number of terms in office did not significantly predict survey response.

Table 4: Knesset Sample Characteristics (N=89)

<table>
<thead>
<tr>
<th>Proportion of respondents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Knesset Member:</td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>25%</td>
</tr>
<tr>
<td>Former</td>
<td>75%</td>
</tr>
<tr>
<td>Exp. on Foreign Affairs/Defense Committee ...</td>
<td></td>
</tr>
<tr>
<td>...as backup or full member</td>
<td>67%</td>
</tr>
<tr>
<td>...as full member</td>
<td>54%</td>
</tr>
<tr>
<td>Highest level of experience:</td>
<td></td>
</tr>
<tr>
<td>...not a Minister</td>
<td>58%</td>
</tr>
<tr>
<td>...Deputy Minister</td>
<td>29%</td>
</tr>
<tr>
<td>...Cabinet Member or higher</td>
<td>12%</td>
</tr>
<tr>
<td>Male</td>
<td>84%</td>
</tr>
<tr>
<td>Served in military</td>
<td>95%</td>
</tr>
<tr>
<td>Active combat experience</td>
<td>64%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>61.4</td>
<td>10.7</td>
</tr>
<tr>
<td>Terms in Knesset</td>
<td>3.0</td>
<td>2.1</td>
</tr>
<tr>
<td>Military Assertiveness</td>
<td>0.61</td>
<td>0.20</td>
</tr>
<tr>
<td>Right Wing Ideology</td>
<td>0.45</td>
<td>0.24</td>
</tr>
<tr>
<td>Hawkishness (Arab-Israeli conflict)</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>International Trust</td>
<td>0.40</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Note: individual differences in bottom four rows scaled from 0-1.
Table 5: Sample representativeness tests

<table>
<thead>
<tr>
<th>Compared to...</th>
<th>All Knesset members</th>
<th>Sampling frame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Current member</td>
<td>−0.043</td>
<td>−0.049</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.057)</td>
</tr>
<tr>
<td>Highest level of experience:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>. . . Deputy minister</td>
<td>0.017</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
<td>(0.071)</td>
</tr>
<tr>
<td>. . . Cabinet member or higher</td>
<td>−0.044</td>
<td>−0.098</td>
</tr>
<tr>
<td></td>
<td>(0.076)</td>
<td>(0.098)</td>
</tr>
<tr>
<td>Male</td>
<td>0.025</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>(0.053)</td>
<td>(0.063)</td>
</tr>
<tr>
<td>Terms in office</td>
<td>0.011</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Left-right party membership</td>
<td>−0.070**</td>
<td>−0.063</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.177***</td>
<td>0.312***</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
<td>(0.087)</td>
</tr>
<tr>
<td>N</td>
<td>415</td>
<td>295</td>
</tr>
<tr>
<td>R²</td>
<td>0.007</td>
<td>0.043</td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01

5.3 Supplementary results
Figure 15: Comparing our downsamled mass public results with those from foreign decision-makers

Figure 15 compares the bootstrapped average treatment effects for three factors from the conjoint experiment, in blue (calculated using B=2500 clustered bootstraps), with the bootstrapped average treatment effects for the same three factors in a pair of survey experiments fielded on an elite sample of members of the Israeli Knesset, in red (calculated using $B = 2500$ bootstraps). The results for regime type are shown in the left-hand panel, and results for costly signals in the middle and right-hand panels. Collectively, they show that the heavier tails in the ATE distributions for the Knesset results in the main paper is due to the relatively small sample size rather than oddities about the sample; when we downsample the public results, we find the distributions have similar spread. For downsampling the public results for regime type in the left-hand panel, we sample $N = 89$ observations with replacement; because of the costly signaling experiment’s between and within-subjects design (in which all participants are also administered the control condition, unlike in the public conjoint experiment), we sample $N = 135$ observations from the public sample with replacement, to ensure that in expectation there are as many sampled observations in each costly signaling condition in the public sample as there would be in its elite counterpart.
5.4 Experiment Instrumentation

5.4.1 Regime Type Experiment

Here is the situation:

- Two countries are currently involved in a public dispute over a contested territory. The dispute has received considerable attention in both countries, because of the risk that disputes like these can escalate to the use of force.
- Country A is a [democracy/dictatorship]. Country B is a dictatorship.
- Both countries have moderately powerful militaries, with large armies, moderate sized navies, and well-trained air forces.
- Neither country is a close ally of the United States.
- Country A is slightly larger than Country B, though their economies are approximately the same size.
- Country A has moderate levels of trade with the international community. Country B has high levels of trade with the international community.
- The last time the two countries were involved in an international dispute, different leaders were in power.

1. Given the information available, what is your best estimate about whether Country A will stand firm in this dispute, ranging from 0% to 100%?

2. If the dispute were to escalate and war were to break out, what is your best estimate about whether Country A will win, ranging from 0% to 100%?

5.4.2 Costly Signals Experiment

Here is the situation:

- Your country — Israel — is involved in a dispute with Country B, a strong military dictatorship.
- The dispute began with a collision between an Israeli shipping vessel and a ship registered to Country B.
• During the collision, injuries were reported on both sides.

• Additionally, both countries maintain that their ship was carrying sensitive military technology, and are suspicious of the motives of the other side, leading to a tense standoff at sea.

• Currently, because of the remote location, the public is not aware of the incident.

[Outcome 1 (Baseline)] Given the information available, what is your best estimate about whether Country B will stand firm in this dispute, ranging from 0% to 100%?

[NEW SCREEN]

Now we would like to ask you a question about a different, alternative version of the scenario you just read. Suppose the basic details remain the same:

• Israel is involved in a dispute with a dictatorship with a strong military, Country B.

• The dispute began with a collision between an Israeli shipping vessel and a ship registered to Country B. During the collision, injuries were reported on both sides.

• Both countries maintain that their ship was carrying sensitive military technology, and are suspicious of the motives of the other side, leading to a tense standoff at sea.

• Currently, because of the remote location, the public is not aware of the incident.

But this time, suppose that . . .

×

[Tying Hands]: The President of Country B has issued a public statement through the news media warning that they will "do whatever it takes" to win this dispute.

[Sinking Costs]: Country B has mobilized their military and sent additional gunboats to the location of the dispute at sea.

Outcome 2 (Treatment)] Given the information available, what is your best estimate about whether Country B will stand firm in this dispute, ranging from 0% to 100%?
References


