Abstract: Do costly signals work? Despite their widespread popularity, both hands-tying and sunk-cost signaling have come under criticism, and there's little direct evidence that leaders understand costly signals the way our models tell us they should. We present evidence from a survey experiment fielded on a unique sample of elite decision-makers from the Israeli Knesset. We find that both types of costly signaling are effective in shaping assessments of resolve, for both leaders and the public. However, although theories of signaling tend to assume homogenous audiences, we show that leaders vary significantly in how credible they perceive signals to be, depending on their foreign policy dispositions, rather than their levels of military or political experience. Our results thus encourage IR scholars to more fully bring heterogeneous recipients into our theories of signaling, and point to the important role of dispositional orientations for the study of leaders.

Acknowledgements: We thank Nick Anderson, Katie Beall, Sarah Bush, Sarah Croco, Matt Fuhrmann, Charlie Glaser, Alexandra Guisinger, Ron Hassner, Mike Horowitz, Tyler Jost, Brad LeVeck, Yon Lupu, Aila Matanock, Michaela Mattes, Bob Powell, Elizabeth Saunders, Rob Schub, Rachel Stein, Caitlin Talmadge, Todd Sechser, Jack Snyder, Mike Tomz, Jane Vaynman, and audiences at GWU, UC Berkeley, UC Merced, Temple University, APSA 2016 and Peace Science 2016 for helpful feedback. This is one of several joint articles by the authors; the ordering of names reflects a principle of rotation.

1Assistant Professor of Politics and International Affairs, Woodrow Wilson School, Princeton University. Email: kyarhi@princeton.edu. Web: http://scholar.princeton.edu/kyarhi/
2Assistant Professor of Government, Harvard University. Email: jkertzer@gov.harvard.edu. Web: http://people.fas.harvard.edu/~jkertzer/
3Assistant Professor, Department of Political Science, University of Wisconsin-Madison. Email: renshon@wisc.edu. Web: http://jonathanrenshon.net
1 Introduction

Much of the “tragedy” of international politics can be reduced to two dynamics that operate in tandem. On one side are leaders trying to accurately assess the capabilities, intentions, and resolve of others, in an international system that incentivizes the weak to appear strong, revisionists to appear status-quo seeking, and the weak-willed to appear resolute (Jervis, 1976; Schweller, 1994; Tang, 2008; Holmes, 2013; Kertzer, 2016). On the other are leaders trying to credibly convey information about their “type” to their allies and adversaries, despite these fundamental incentives to misrepresent (Jervis, 1970; Fearon, 1995; Sartori, 2002; Trager, 2010). According to an important tradition in IR, the solutions to these challenges are costly signals: messages, gestures, or actions that are costly enough that only an actor of a certain type would be able, or willing, to them carry out (Schelling, 1960; Fearon, 1997; Jervis, 2002; Kydd, 2005; Slantchev, 2005; Fuhrmann and Sechser, 2014). Issuing a threat in public, for example, is a costly signal if audiences — either at home or abroad — are in position to impose costs on leaders who back down, such that leaders should shy away from making empty threats (Fearon, 1994; Tomz, 2007; Weeks, 2008; Kertzer and Brutger, 2016). Mobilizing troops is another type of costly signal, in that only a leader who thought an issue was truly worth fighting for would be willing to pay those kinds of costs upfront (Fearon, 1997; Sechser and Post, 2015).

A barrage of scholarship, however, suggests that IR scholars’ faith in costly signals may be misplaced. Political psychologists lament that the study of signaling has been largely divorced from the study of perception, such that many of our models of signaling assume too much: signals are often misinterpreted and messages lost in translation (Jervis, 1976, 2002; Mercer, 2012; Grynaviski, 2014). Another line of critique argues focuses less on the deficiencies of the recipient, and more on the weaknesses of the particular types of signals IR scholars tend to study: two decades ago, IR scholars suggested military mobilization was relatively uninformative compared to tying hands (Fearon, 1997); now, public threats are under assault as well (Snyder and Borghard, 2011; Trachtenberg, 2012; Levendusky and Horowitz, 2012). On top of these substantive critiques, formal theorists and experimentalists have raised methodological concerns, warning that — as a result of strategic behavior — the effects of some types of signals (such as tying one’s hands by making public threats) might be difficult to uncover in the historical record (e.g., Schultz, 2001; Tingley and Walter, 2011a). Thus, despite the accumulation of research on costly signaling, a skeptic can be forgiven for asking:
do costly signals work, after all? Do leaders understand costly signals the way our models in IR tell us they should?

In this piece, we aim to make two contributions. First, we turn to survey experiments to test the microfoundations of costly signaling. Unlike other experimental work on signaling, however, we test these theories on a sample of past and present political leaders, the exact population around whom many of our theories are built. Our participants, drawn from the Israeli Knesset, are not only elite in every sense of the term (ranking all the way up to Prime Minister), but also have a history of foreign policy decision-making, with over two-thirds of our sample having had experience on the Foreign Affairs and Defense Committee, giving us vital insight as to how elite foreign policy leaders interpret costly signals. Second, although the signaling literature tends to assume that all recipients should process the same signal in the same way, we build on a growing body of interest in the ways in which leader-level characteristics and experiences shape decisions about war and peace (e.g., Horowitz, Stam and Ellis, 2015; Yarhi-Milo, 2014; Fuhrmann and Horowitz, 2015; Kertzer, 2016; Saunders, 2017, Paper #3 This Issue, Paper #5 This Issue) by exploring whether leaders with different types of military and political experiences and foreign policy orientations perceive signals differently.

Our main findings are fourfold. First, we show that costly signaling is effective in updating leaders’ assessments of resolve; in a parallel experiment on a nationally representative sample of Israeli Jews, we find that the mass public draws these inferences as well. Second, in both samples, we show that sinking costs are no less persuasive to receivers than tying hands, in contrast to Quek’s (2016) recent work finding that sinking costs did little to influence players’ decisions in a bargaining game. Third, in the public sample, we fail to find evidence consistent with claims of democratic signaling advantages (Fearon, 1994; Schultz, 1999; Downes and Secher, 2012): public threats are effective in the eyes of foreign observers, but their credibility is not conditional upon the regime type of the sender. Finally, we show that, against the assumption of recipient homogeneity, different types of leaders calculate the credibility of signals differently. Interestingly, these differences have more to do with general dispositions towards foreign affairs, rather than levels of military or political experience, and in a manner that suggests that tying hands and sinking costs are imperfect substitutes for one another, since hawks and doves perceive signals in different ways. The results remind us to bring theories of the recipient back into theories of signaling and reinforce that — although leader-level characteristics matter — IR scholars shouldn’t restrict the study of these
characteristics to experiential variables alone.

In the discussion that follows, we begin by reviewing the costly signaling work in IR, showing how, as models of costly signaling have grown in popularity, so too have doubts about whether these signals work the way our models tell us they should. We then turn to the growth of interest in leader characteristics, which has tended to argue that experiences shape leaders’ beliefs about the world, which then affect how they interpret information and make decisions, thereby linking the study of signaling with the study of leaders. Subsequently, we present our experimental designs before presenting our experimental results. We conclude by discussing the implications of our findings for the study of signaling, and the study of leaders in IR more generally.

2 Leaders and Signaling

One of the most significant problems in international politics is how states can learn about the resolve of their adversaries. Unlike tangible attributes like military strength or wealth, resolve is typically understood to be private information that decision-makers can access but foreign rivals cannot (though see Kertzer, 2016, 148-149). In his seminal work on crisis bargaining, Fearon (1997) describes two types of costly signals states can use to persuade others of their resolve. The first is through mobilizing troops, a signal in which the actor pays the costs upfront (or, ex ante) regardless of the eventual outcome. Military mobilization is thus a “sunk cost,” and because only actors who value an issue sufficiently would be willing to pay these costs, it acts as a credible signal of resolve. The second is through issuing public threats, which generates political (or “audience”) costs that the actor will pay ex post should they fail to follow through. The ability for leaders to tie their hands by making threats in public thus allows them to signal resolve to adversaries; even though they pay the political costs only if they back down, their willingness to risk escalation acts as a credible (and costly) signal of resolve. While these tactics differ in when and how costs are paid (and whether they are conditional on behaviors or outcomes), both are intended to manipulate the beliefs of others. They represent a method by which actors may solve a problem known as type separation; if there are high costs to bluffing, then only truly resolved actors would make public threats, allowing adversaries to distinguish resolute and irresolute states.

Both of these models of signaling have become enormously influential in IR. Scholars have ap-
plied sunk cost models to everything from the study of covert interventions (Carson and Yarhi-Milo, 2016) to reassurance (Glaser, 1994; Kydd, 2005) to military alliances (Morrow, 1994). Similarly, the study of public threats has led to an explosion of research on audience costs, whether game theoretic (Fearon, 1994; Schultz, 2001), quantitative (Weeks, 2008; Potter and Baum, 2010), qualitative (Snyder and Borghard, 2011; Trachtenberg, 2012; Weiss, 2013), or experimental (Tomz, 2007; Trager and Vavreck, 2011; Chaudoin, 2014; Kertzer and Brutger, 2016). Yet a survey of the theoretical and empirical landscape suggests a degree of caution.

First, skeptics have questioned both the logic and efficacy of mobilization. For example, Slantchev (2005, 2010) argues that — contra the logic proposed by Fearon (1997) — mobilization both sinks costs (i.e., states pay the costs regardless of what follows) and ties hands (by increasing the odds of winning should war occur). In Slantchev’s story, military mobilization is an effective signal of resolve, but acts simultaneously through two mechanisms: incurring costs (to separate low- from high-resolve types) and by changing the expected value of war by improving one’s prospect for victory. Empirical concerns abound as well. Quek (2016), using economics-style bargaining games, found that while resolved signalers were more likely to sink costs, receivers did not acquiesce in line with expectations. Meanwhile, the empirical record is riddled with examples of leaders mobilizing troops or stationing nuclear weapons in order to demonstrate resolve, and yet, Fuhrmann and Sechser (2014) despite finding evidence for the ability of states to tie their hands, found no evidence that this “sinking costs” mechanism deterred adversaries as expected (one of the few empirical efforts to directly compare the effects of these two types of signals head-to-head).

Critics have expressed similar grievances about the logic and efficacy of public threats. Initially, the consensus was that the ability to use public threats for hands-tying offer democratic leaders an advantage in crisis bargaining. Many scholars have found that public opinion seems to broadly follow the logic of audience costs (Tomz, 2007; Trager and Vavreck, 2011; Davies and Johns, 2013), although some report a more complex picture, finding that audience costs are not imposed exactly as audience cost models seem to suggest (Levendusky and Horowitz, 2012; Chaudoin, 2014; Kertzer and Brutger, 2016). For example, Weeks (2008) showed that the ability to signal resolve through this mechanism was not limited to democracies. Empirically, the efficacy of hands-tying in signaling resolve is unclear. Using historical case studies of international crises, Snyder and Borghard (2011) find that leaders rarely issue clear public threats because they view them as imprudent; domestic audiences

---

See also Weiss 2013 on how autocrats can use nationalist protests to generate costly signals.
appear not to care about a leader’s consistency between words and deeds relative to concerns about policy substance and reputation. Moreover, they find little evidence that authoritarian targets of democratic threats perceive audience cost dynamics in the same way that political scientists expect them to. Similarly, Trachtenberg (2012, 7) argues that “the audience costs mechanism can be decisive only if the opposing power understands why it would be hard for those leaders to back down. Unless the adversary is able to see why the democratic power’s leaders’ hands are tied, it would have no reason to conclude that they are not bluffing.” And yet, Trachtenberg finds that even during major international crises, leaders rarely took the American domestic political context into account.3

There are at least three clusters of reasons why the empirical record for these costly signaling mechanisms are so mixed, with some scholarly work finding solid evidence in their favor (Lai, 2004) and others suggesting caution or revision to our theoretical frameworks. The first has to do with questions about the logic of the signaling models themselves. Snyder and Borghard (2011), for example, find that leaders rarely pay substantial audience costs, perhaps because domestic audiences “care more about policy substance than about consistency between the leader’s words and deeds.” Levendusky and Horowitz (2012) similarly show that while signals matter, other factors impact how signals are perceived and interpreted. Presidents can, for example, justify backing down based on new information and earn themselves a substantial discount on the “audience cost” they would otherwise pay. More generally, though, since signaling models inherently involve uncertainty about the signalers’ type — without it, actors would have no reason to signal in the first place — the null effects across many of these studies could also simply reflect this underlying uncertainty rather than the illogic of the signals themselves.

The second concerns the analytic difficulty of identifying the effects of signals in a strategic environment. It is possible, for example, that the datasets IR scholars typically use may be inappropriate for questions relating to signaling. Downes and Secher (2012) show that the Militarized Interstate Disputes (MID) and International Crisis Behavior (ICB) datasets that have typically been used to test audience cost theory contain very few of the coercive threats that are necessary to test it. More generally, we may be misguided in looking for observational evidence to begin with. As summarized by Dafoe, Renshon and Huth (2014, 385; see also Schultz 2001): “[i]f leaders care about reputation (or domestic support), we can expected the observed effects [of backing down] to be biased towards

3In this sense, much of the skepticism about costly signaling in IR is less about whether signals matter, per se, and more about the efficacy of the particular types of signals that IR scholars have tended to study.
The third reflects potential mismatches between signals’ senders and receivers. Most theories of costly signals leave little room for interpretation: a signal of magnitude $m$ is sent by $i$, and thus a signal of magnitude $m$ is received by actor $j$. Using samples of college students and American participants on Amazon Mechanical Turk, Quek (2016) found that sunk costs did not have the effect of deterring other actors in line with expectations, suggesting that either $j$ did not receive the signal as intended, or that the message was received but did not affect their behavior.

This type of “sender-receiver gap” can be attributed to two sources: systematic biases in how individuals process information, or the effects of leaders who come to office with different experiences, beliefs and orientations. The first of these harks back to a classic literature in political psychology concerning systematic and predictable biases in how actors perceive and interpret signals. Mountains of research in political, cognitive and social psychology have demonstrated that signals are filtered through ideologies and belief systems (Herrmann, Tetlock and Visser, 1999), enemy images (Herrmann and Fischerkeller, 1995), analogies (Khong, 1992), and emotions (Mercer, 2010; Hall and Yarhi-Milo, 2012). Thus, our models and theories might need updating to account for how actors typically update beliefs about resolve.

A second, related issue concerns the characteristics of $i$ and $j$. Our theories of signaling tend to treat all actors as the same — both for the purposes of elegant modeling, and because without more fine-grained theories and data it would have been foolhardy to begin with heterogeneous types in our models. However, a growing literature in IR has demonstrated the importance of experiences and beliefs in accounting for leaders’ behavior (Goldgeier, 1994; Kennedy, 2012; Khong, 1992). For example, Horowitz and Stam (2014) has demonstrated the importance not just of military service, but of particular types of combat in affecting leaders’ behavior in world politics. If leaders systematically differ from one another on a range of attributes (Kertzer, 2016), and these differences affect how they interpret information and define the situations they face, signals may fail because of a mismatch between the expectations of the signaler and the interpretations of the recipient.

This brief overview of the literature leaves us with three sets of questions relating to signaling theory in international relations. First, do costly signals work? Second, are some types of costly signals more effective than others in sending credible signals of resolve? And finally, do the characteristics — such as their experiences or orientations — of the recipient matter in explaining variations in the signals’ efficacy? Below we briefly discuss our methodological approach and hypotheses.
2.1 Hypotheses

Of the three broad questions that are the focus of this paper, the first presents a relatively stark, clear hypothesis. If a rationalist approach accurately captures how costly signals shape assessments of credibility, leaders’ estimates of adversaries’ resolve should increase when costly signals (of any type) are utilized (H1). The second question leads us into slightly murkier waters. As we noted, there is a vibrant theoretical debate, though little empirical consensus, on which types of signals might be (relatively) more effective. Thus, our second set of hypotheses (H2) concerns whether public threats constitute a more, less, or equally credible signal of resolve compared to military mobilization.

For many years, scholars wishing to study individual-level heterogeneity turned to a literature on cognitive and motivational psychological bias (the “heuristics and biases” approach - see Kahneman and Tversky 2000). For example, the difficulty of updating beliefs suggests not so much an additional set of hypotheses, but rather the null hypothesis for H1 and H2. A wide body of findings suggest that these beliefs are updated sporadically, and then only in response to large events, or when leaders are “shaken and shattered into doing so” (Stoessinger 1981, 240; see also Jervis 1976). Thus, if leaders do indeed find it as difficult to change their beliefs about others’ resolve as these works suggest, we should expect either an attenuated effect of signaling or none at all.

In contrast to a focus on biases, we approach the issue of leader-level heterogeneity through a focus on orientations and experiences. In doing so, we turn to a burgeoning literature in international politics on the relationship between leaders’ characteristics and international conflict.

Leaders’ orientations provide the first lens through which to understand leaders’ behavior and credibility judgments (H3). Although there are a wide array of orientations one could choose from, for reasons of tractability we focus on two here. First is military assertiveness, the ubiquitous division in IR between hawks and doves that has been shown to significantly shape both attitudes and behavior in foreign policy crises (Snyder and Diesing, 1977; Herrmann, Tetlock and Visser, 1999). Hawks and doves differ from one another in two interrelated respects (Glaser, 1992; Brutger and Kertzer, 2017). The first concerns their beliefs about the nature of adversaries in the international system, with hawks subscribing to a “deterrence mindset” that understands adversaries as being driven by expansionist goals, and doves embracing a “spiral model” that instead attributes conflict to misperceptions (Jervis, 1976). The second concerns their beliefs about the desirability and efficacy
H1: Leaders will view the adversary’s public threats to escalate the crisis or the mobilization of troops during a crisis as a credible signal of the adversary’s resolve, and will accordingly update their beliefs about the adversary’s likelihood of standing firm.

H2a: Leaders will view the adversary’s public threat to escalate the conflict as a more credible signal of resolve compared to the adversary’s mobilization of troops.
H2b: Leaders will view the adversary’s public threat to escalate the conflict as equally or less credible signal of resolve compared to the adversary’s mobilization of troops.

H3a (Orientations: Military assertiveness) Leaders who are hawks will view military mobilization as a more informative signal of resolve than leaders who are doves, but be less persuaded by mere verbal threats.
H3b (Orientations: International trust) Leaders who are higher in international trust should view mobilization and public threats as more credible indicators of intentions compared to leaders who are low in international trust.

H4a (Experience: Bayesian Updaters) Leaders with more experience should view costly signals as more informative signals of resolve than leaders with less experience.
H4b (Experience: Biased Learners) Leaders with more experience should be less likely to update from costly signals of resolve than leaders with less experience.

Figure 1: Hypotheses
of the use of military force to achieve their foreign policy objectives. Given the extent to which hawks view international politics through the prism of force, we expect leaders high in military assertiveness to view military mobilization as a significantly more credible signal of resolve than doves do. Because doves embrace the spiral model, they should perceive military gestures as inherently ambiguous, and thus perceive them less credible signals of actors’ intentions. In comparison to military mobilization, hawks should be relatively skeptical about assigning credibility based on public threats, since they do not directly involve the use of military force, and are thus more likely to be discounted as mere “cheap talk” (H3a).

In addition to military assertiveness, scholars have highlighted the important role that trust plays in shaping leaders’ beliefs, judgments and attitudes toward conflict and cooperation (Larson, 1997; Brewer et al., 2004). Trust, as typically conceptualized in the literature, is dispositional rather than a situational factor determined by players’ incentive structures: setting aside the incentives to trust others present in any given situation, some actors are inherently more trusting than others (Rathbun, 2011). These individuals are thus more likely to cooperate, and less concerned that their trust will be exploited by others. This insight has important implications for individuals’ attitudes towards international crises. For example, Kertzer and Brutger (2016) show that because individuals low in international trust tend to assume others are likely to exploit them, they tend to make judgments in accordance with the predictions of audience cost theories by punishing leaders who say one thing and do another. In our context, individuals high in international trust should be more likely to take other nations at their word, and less concerned at the thought of being exploited, misled or lied to. Mobilization and public threats are fundamentally signals that are designed to convey one’s intended behavior in the crisis. Therefore, those higher in international trust should perceive both types of signals as credible indicators of how the other country will behave if one does not back down, and thus should assign more credibility to them than individuals who are lower in international trust (H3b).

Whereas dispositional orientations offers one important lens through which leaders process information, another optic frequently studied in the IR literature (and in the other articles in this issue: Paper #5) is leader experiences. It has long been established that individuals’ experiences shape their dispositions, inclinations and patterns of behavior. When evaluating the relationship between experience and international conflict more specifically, scholars have highlighted the role of two types of experience: military and political.
Earlier work has come to varying conclusions about how military experience should affect the behavior of leaders in crisis (Huntington, 1957; Sechser, 2004; Feaver and Gelpi, 2005). Recently, Horowitz and Stam (2014) found that leaders with combat experience who rise to power in democracies tend to react to their experience in combat dispassionately, reducing their inclination to use military force (in contrast to autocratic leaders with combat experience). Scholars have also long probed the relationship between leaders’ political experience and their international conflict behavior. Most of these studies (Bak and Palmer, 2010; Gelpi and Grieco, 2001) use leader age and tenure in office as proxies for experience. Calin and Prins (2015), in contrast, look at the role played by the president’s past executive experience in determining foreign policy behavior. They offer quantitative evidence suggesting that the higher the level of a president’s executive experience prior to assuming office, the less likely the U.S. was to be involved in conflict (either as a target or initiator) during their tenure. Saunders (2017) argues that leaders with experience can better monitor advisers, more effectively delegate, and obtain diverse advice. She finds that George W. Bush’s inexperience exacerbated the biases of his advisers during the 2003 Iraq war, whereas his father’s experience cast a long shadow over many of the same officials during the 2001 Iraq war. When experience is measured using leader age, it appears that older leaders are more likely to participate in militarized conflict and that leaders who have been in office longer have a higher propensity for conflict involvement (Calin and Prins, 2015).

Experience not only affects leaders’ propensity to use military force, but also their judgment and information processing, though in somewhat unpredictable ways. On the one hand, Hafner-Burton, Hughes and Victor (2013, 369) suggest that experienced elites tend to act more strategically and cooperatively, display greater reliance on the “right” heuristics, and play iterated games more effectively. They conclude that decision-making by experienced elites “more closely approximates the canonical rational actor assumption” that also underlies the costly signaling framework (on rationality as a variable, see Rathbun, Kertzer and Paradis, 2017). If experienced leaders act more like rational actors, they are likely to view costly signals as diagnostic information and will use this to update their beliefs in response to costly signals to a greater extent than inexperienced leaders (H4a). These studies thus suggest that we should expect experienced leaders to be more attuned to costly signals that reveal new information about the adversary’s resolve during international crises.

Other studies have shown that experience and expertise do not turn leaders into good Bayesian updaters, but instead exacerbate biases. For example, Tetlock’s (2005) work on experts reveals
that contrary to the conventional wisdom, experts' judgments and predictions are no better than novices on many important political questions. Experts in his studies were especially susceptible to confirmation bias, more likely to use inappropriate heuristics, and more reluctant to update beliefs in response to new information. Power, Tetlock contends, tends to exacerbate the overconfidence and risk-taking of experts, suggesting that when experienced leaders are in power, they might deviate even more from the expectations of rational choice. Moreover, experienced leaders have been found to rely on their own intuition and pre-existing beliefs to gauge the adversary's intentions in crisis more so than the adversary's crisis signals, leading them to fail to update as rationalist accounts would predict (Yarhi-Milo, 2014; Tetlock, 2005). If experienced leaders are more prone to confirmation and other related types of biases, we would expect them to be no more responsive to costly signals during international crises compared to inexperienced leaders; and they might even exhibit more resistance to updating their beliefs in response to costly signals than inexperienced leaders (H4b).

3 Research Design

Earlier, we noted several methodological problems — primarily related to the use of observational data — that have slowed the development of a cohesive body of evidence. However, another significant issue has been the reliance on behavioral proxies for credibility, since theories of costly signaling ultimately hinge on the content of leaders' beliefs or perceptions. To address these issues in tandem, we turn to experimental methods. Experiments provide significant benefits in making causal inferences, which are by now sufficiently well known as to not require repeating (McDermott, 2002; Hyde, 2015), but we add two twists to our experimental design that innovate and build upon previous work in this area. First, in contrast to some studies (e.g., Tingley and Walter, 2011b; Quek, 2016), we center our costly signaling experiment not on an economics-style bargaining game, but an IR scenario that reflects the substance of what our theories of interstate costly signaling are designed to explain. Second, in contrast to much other experimental work in IR, we utilize a unique sample of real-world elites — former and current members of the Israeli Knesset — rather than college students. While all experimental work requires serious consideration of how the results might extrapolate to other populations, domains, and so on, ours comes unusually close to directly studying the population of interest.

Studying Israeli decision-makers offers a number of substantive advantages. First, if our goal
is to examine decision-makers’ beliefs about resolve in international crises — the concept on which our theories hinge — then the most direct way to do so is to sample from exactly that population: leaders who have had to wrestle with these issues outside the lab. This is, after all, what is unique about studying leaders rather than the general public. In fact, our elite participants have repeatedly encountered these issues, as “use of force” decisions have been ubiquitous over the past few decades in Israel: during the time frame in which our sample of leaders were in office (from 1996 onwards), Israel was involved in 16 Militarized Interstate Disputes — and we can reasonably expect our subjects to have actually been involved in the decision making process for many of these cases.

Second, because of the structure of Israel’s parliamentary system (most of the executive branch are also elected members of parliament), the Israeli Knesset — unlike, e.g., the U.S. Congress — is comprised of policy makers who are directly involved in use of force decisions. And as a result of political norms and relatively short election cycles, it is common for former members of the executive branch (i.e., Ministers and Prime Ministers) to later become members of the opposition in the Knesset; conversely, nearly all current members of the executive branch were at some point in their career members of the opposition in the Knesset. Thus, even the Knesset members in our sample who are currently part of the opposition have either been members of the executive branch in the past, or are likely candidates to become members in the future.

In addition to directly examining the beliefs of Israeli leaders, we field a followup study on a representative sample of the Jewish public in Israel. This pairing — elite leaders and the general population — provides several important benefits. The first and most obvious is the value inherent in any replication, which increases our confidence in the validity of the individual study as well as the overall research program. This also allows us to gain some leverage on the selection of leaders from the general population, thereby giving us insight into the dimensions on which they differ and those which they resemble their compatriots. Methodologically, the paired experiments allow us to more properly gauge the utility of experimental IR studies carried out on normal citizens that are — for logistical reasons — likely to remain the norm for the foreseeable future.

As with the Israeli leaders, there are reasons to suspect that Israeli citizens might be particularly suitable for our purposes. For example, the direct impact of foreign policy on the daily lives of Israelis suggests that they are likely to be far more informed of foreign policy issues than is the case in most

---

4 Our focus on the Israeli Jewish population alone is due to logistical constraints posed by the lack of existing internet-based polling companies in Israel providing anything close to a representative sample of the minority Israeli Arab population.
representative democracies. Additionally, since all Israeli citizens are required by law to enlist in the military at the age of eighteen, questions about foreign policy and the use of force are extremely salient for members of the Israeli public. The result of this is that our Israeli citizens should be more likely to be accessing actual beliefs about the use of force in foreign policy (rather than constructing belief systems “on the fly”) as well as paying closer attention to the experimental vignette (increasing the validity of their responses). As in any question about case selection, the choice of Israel also raises a number of questions about generalizability, which we consider in Appendix §3.2.

3.1 Experimental Samples

The Knesset Subjects

We fielded the study from July - October 2015. Of 288 potential subjects, 89 participated, leaving us with a 31% response rate, relatively high for surveys of this type. The Knesset sample is described in Table 1 and our recruitment procedures are described in Appendix §1. 25% of our participants were currently serving in the Knesset; the rest of the sample (75%) was composed of former Knesset members. They were also highly experienced in IR-relevant contexts: 64% had active combat experience, and 67% had experience serving as members of the Knesset’s Foreign Affairs and Defense committee. In addition to experience along the dimension of military strategy, they also had considerable political experience: our participants served an average of 3 terms in Parliament (with some as high as 9 terms). While 61% of the Knesset subjects had never served as a Minister, 29% had been at least a Deputy Minister, and fully 12% of our sample was in our highest category of elite experience, such that our participants include individuals who had served as Cabinet members, and even Prime Minister.

Given the intricate nature of recruiting elite participants for social science research, it is worth considering how representative our leader sample is of the universe of Israeli political leaders from the time frame we examined. This requires thinking through how our subjects compare both to the overall universe of Knesset members from 1996 onwards, as well as the sampling frame (which does not include members who passed away, were too sick to participate or for whom we could not obtain contact information). In Appendix §3.1, we conduct supplementary analyses showing that our sample is fairly representative, although not surprisingly, current Knesset members were less

---

5Men serve a period of three years and women two years. In practice, however, Israeli Arabs, ultra-Orthodox Jews, and religious women are exempt from serving in the IDF.
Table 1: Knesset Sample (N = 89)

<table>
<thead>
<tr>
<th>Proportion of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Knesset Member:</strong></td>
</tr>
<tr>
<td>Current</td>
</tr>
<tr>
<td>Former</td>
</tr>
<tr>
<td><strong>Exp. on Foreign Affairs/Defense Committee . . .</strong></td>
</tr>
<tr>
<td>. . . as backup or full member</td>
</tr>
<tr>
<td>. . . as full member</td>
</tr>
<tr>
<td><strong>Highest level of experience:</strong></td>
</tr>
<tr>
<td>. . . not a Minister</td>
</tr>
<tr>
<td>. . . Deputy Minister</td>
</tr>
<tr>
<td>. . . Cabinet Member or higher</td>
</tr>
<tr>
<td><strong>Male</strong></td>
</tr>
<tr>
<td><strong>Served in military</strong></td>
</tr>
<tr>
<td><strong>Active combat experience</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>61</td>
</tr>
<tr>
<td>Terms in Knesset</td>
<td>3.0</td>
</tr>
<tr>
<td>Military Assertiveness</td>
<td>0.61</td>
</tr>
<tr>
<td>Right Wing Ideology</td>
<td>0.45</td>
</tr>
<tr>
<td>Hawkishness (Arab-Israeli conflict)</td>
<td>0.39</td>
</tr>
<tr>
<td>International Trust</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Note: individual differences in bottom four rows scaled from 0-1.

likely to participate than former members.

Studies of “elites” in political science have grown more common in recent years. Renshon (2015) uses political and military leaders drawn from a mid-career training program at Harvard Kennedy School, while Alatas et al. (2009) use Indonesian civil servants, Hafner-Burton et al. (2014) use “policy elites” (including civil servants, senior executives in international firms, former members of Congress, and U.S. trade negotiators) and Mintz et al. (1997) use Air Force officers. While more “elite” than college freshmen, to be sure, the samples used are still somewhat removed from the dictators, presidents, leaders of the military and foreign ministry, trusted advisors, and generals who are the primary decision makers in most interstate conflicts. This serves as a reminder that, even among the still relatively rare genre of elite experiments, our subjects are unusually close to the levers of power in their country.6

6Recent work in a similar style includes Loewen et al. (2014); Linde and Vis (2016).
The Israeli Public

Our followup study used a representative sample of the Israeli general public obtained by iPanel, an Israeli polling firm that has been used effectively by other recent surveys and experiments (e.g., Ben-Nun Bloom, Arikan and Courtemanche, 2015). The sample was representative of the Israeli Jewish population, and stratified based upon gender, age, living area and education. This study was fielded in autumn 2015. Descriptive statistics, and a comparison of the public and Israeli samples on a variety of covariates, can be found in Appendix §4.

3.2 The Experiment

Figure 2: Experimental design

I. Baseline scenario

1. Describe dispute scenario

II. Costly signal treatments

2. Measure assessment of Country B’s resolve

3. Tying hands signal

3. Sinking costs signal

4. Measure assessment of Country B’s resolve

Note: in the leader survey, Country B is a dictatorship; in the public survey, the regime type of Country B is randomly assigned (either dictatorship, or democracy).

Our experiment – the structure of which is depicted in Figure 2 – was fielded first on Israeli leaders and then replicated on a representative sample of the Israeli public. The experiment presented subjects with a vignette that described a dispute between Israel and another country and asked subjects to estimate the odds that the other country (“Country B”) would stand firm in the dispute. After that outcome question, which functioned as each subjects’ 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. Thus, our study combines 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.

7Our focus on the Israeli Jewish population alone in this survey is due entirely to logistical constraints, specifically the inability of online polling companies in Israel to provide anything close to a representative sample of the minority Israeli Arab population.
scenario.

Our experiments were programmed online using Qualtrics (aside from the few Knesset members who chose to fill out paper copies) and presented in Hebrew (but described here in English). The vignette began with a now-standard introduction (“We are going to describe a situation the international community could face in the future. . .”), and a vignette, the full text of which is reproduced in Appendix §2. Note that because of the within-subject design, all subjects read the same initial scenario.

The scenario described the participant’s country, Israel, as being in a dispute with Country B, described as a “dictatorship with a strong military.” In the Knesset study, the regime was fixed, while in the Public study, we randomly assigned subjects to one of two conditions, DICTATORSHIP or DEMOCRACY (both were described as having strong militaries). The vignette then describes a standoff at sea following a collision between Israel and Country B’s ships; in that collision, injuries were reported on both sides, and both countries claim that their side’s ship was carrying sensitive military technology. After telling subject that, because of the remote location, neither country’s public is aware of the “tense standoff,” the instrument asks subjects to estimate the likelihood of Country B standing firm in this dispute.

Following their response (their baseline estimate of B’s resolve), all subjects read another section of text which asked them to think about a different version of the scenario they had just read. In this subsequent scenario, the basic details remained exactly the same but 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.” After this, we again asked subjects to estimate the likelihood of Country B standing firm in the dispute. We therefore use a within-subject design, both to boost statistical power given the necessarily small samples associated with elite experiments, and because it lets us study how participants update, central to the study of signaling.

After the experimental components of the study, subjects completed a demographic and dispositional questionnaire (reproduced in Appendix §2.2).
4 Leaders and Costly Signals

4.1 Do Leaders update beliefs about resolve in response to signals?

Our first task is to evaluate H1, which describes the observable implications that flow directly from the rationalist IR literature. To the extent that there is a theme that ties together this body of literature, it is the notion that leaders should update beliefs about the resolve of other actors when those actors take the specific actions of either escalating the crisis or mobilizing troops. That is, both tying one’s hands and sinking costs should serve as costly signals of resolve.

Given our research design, assessing H1 requires comparing each participant’s baseline assessment of Country B’s resolve to that same participant’s assessment of B’s resolve after being assigned to either the public threat or mobilization treatment: subtracting the former from the latter gives us our (within-subject) treatment effect. The top panel of Figure 3 plots the average treatment effects for the leaders in the Knesset sample. We find strong support for the notion that leaders update their estimates of opponent’s resolve when costly signals are sent (H1). Leaders saw public threats as 8.1% — and mobilization as 6.8% — more credible than the baseline condition. Though the mobilization effect is slightly weaker than that of a public threat, the difference between the two treatments was not itself significant. Thus, in considering H2, we find no evidence that one type of signal is interpreted as being significantly weaker than the other all other factors equal.

The bottom panel of Figure 3 depicts the results from our second study, on a representative sample of the Jewish public in Israel. Most importantly, we replicate our initial finding: it is not just leaders who update beliefs about resolve based upon costly signals. In fact, we find strong and significant effects for our public sample, identical in direction to the results from the Israeli leaders. To the extent that there are differences between the leaders and public, it is a matter of magnitude: the Israeli public saw public threats and mobilization as 22.0% and 20.9% more credible than the baseline condition, more than doubling the size of the effect among leaders.8

As described earlier, our replication study on the Israeli public contained an additional manipulation: we randomly assigned subjects to conditions in which Country B was described as either a Democracy or a Dictatorship (in the leader study, it was fixed at “Dictatorship”). As a result, we are able to offer some insight on the question of democratic credibility. Against theories that

---

8The reported results above are for the entire sample. If we screen out those member of the public who failed our manipulation check, the magnitude of the effect increases slightly (to 24.5% and 22.4%, respectively). For this section of the analysis, we average across regime types (Democracy and Dictatorship) for the public sample.
Figure 3: Effect of costly signals

(a) Leaders

(b) Public

Note: Panel a presents the bootstrapped distribution of average treatment effects in the leader sample; panel b presents the bootstrapped distribution of average treatment effects in the public sample. In both cases, the costly signals are significant and positive, although the effects are larger in the latter. In the public sample, we also manipulate the regime type of the opponent (either a democracy, depicted in light grey, or a dictatorship, depicted in dark grey). Against theories of democratic credibility, it appears that public threats from dictators are seen as more credible than public threats from democracies, although the difference is not statistically significant.
hold that democracies are better able to credibly signal their resolve with public threats than their non-democratic counterparts, but in accordance with some other recent work (Tomz, 2009; Ren- shon, Yarhi-Milo and Kertzer, 2016), we actually find that foreign audiences perceive these actions as a less persuasive signal of resolve when taken by a democracy. For example, public threats issued by democracies increased estimates of resolve by 20.8%, 2.3% less than the same signal sent by a dictatorship, although this difference-in-difference is not statistically significant (bootstrapped $p < 0.175$).\footnote{When we screen out participants who failed the manipulation checks, the difference-in-difference increases in magnitude to −3.15%, although the effect is still not significant (bootstrapped $p < 0.145$).}

### 4.2 How do experiences and beliefs shape how leaders interpret signals?

Having shown that elite political leaders do update their estimates of others’ resolve in response to costly signals, we now turn to the question of how the characteristics of the leaders themselves affect their calculations of credibility. While there are many ways to characterize the dimensions that distinguish leaders from one another, we do so through a focus on experience and orientations, as depicted in Figure 4. The list of experiences studied include both the types of military experiences that have been found to explain variation in leader conflict behavior (e.g., whether leaders served in the military without experiencing combat; Horowitz, Stam and Ellis, 2015), as well as different measures of political experience, whether direct (e.g., experience on the Foreign Affairs committee (e.g. Saunders, 2017, or as rebels (Paper # 5 This Issue)) or indirect (e.g., leaders’ age (Bak and Palmer, 2010))).\footnote{Because of the ubiquitousness of military service in Israel, 95% of our leader sample served in the military, so we are only able to test the effects of combat experience (or the effects of military service without combat), rather than the effects of military service by itself.}

To analyze how leaders’ characteristics affected their estimates of resolve, we estimate a series of regression models, regressing each characteristic on the within-subject treatment effect to test whether leader-level characteristics explain variation in the perceived credibility of each type of signal.\footnote{Because of the mixed experimental design, we split the sample based on the type of costly signal to which participants were randomly assigned (public threat, or military mobilization), and estimate separate regression models for each one. Note that since the treatment effect is measured at the within-subject level (such that we have individual-level treatment effects, rather than just an average treatment effect), we can model treatment heterogeneity by directly regressing the characteristics on the treatment effect, rather than needing to interact the characteristics with a dichotomous variable representing the treatment status. Thus, each of the coefficient estimates should be interpreted as a moderator of the treatment effect under consideration, rather than as the main effect of the particular characteristic on assessments of resolve. Since we are exploiting natural variation in our elites’ experiences and orientations rather than manipulating them, in the discussions below we avoid making causal claims about their effects.} Because of the small sample size, and intercorrelation between a number of our experiential variables, we begin by estimating separate models for each measure of experience, before a full model
Experience

- Combat experience: whether leader has been in combat
- Military experience: whether leader has been in served in military (but not been in combat)
- Terms in office: no. of terms in Knesset
- Foreign Affairs experience: served on Foreign Affairs committee (as a full or backup member)
- Elite experience: highest level of political experience: member of parliament/Deputy Minister/Cabinet Member or higher
- Age: age in years

Orientations

- Military assertiveness: military assertiveness, scaled from 0-1
- International trust: international trust, scaled from 0-1

Figure 4: Leader Characteristics

with all the covariates.\textsuperscript{12} Our main findings are illustrated in Figure 5, which plots the distribution of bootstrapped coefficient estimates for each type of experience and orientation by treatment condition (Mobilization or Public threat). Unlike in economics-style games where payoff structures are controlled by the experimenter, it makes little sense to claim an objectively “accurate” baseline from which to judge whether leaders are perceiving the signals correctly or not; we focus instead on which types of leaders perceive each type of signal as more or less credible (Jervis, 1976).

4.2.1 Leader orientations

We begin with the role of leader orientations, the results of which are presented in the bottom portion of each panel of Figure 5.

Panel (a) on the left shows that hawks see talk as cheap, whereas doves take public threats more seriously. As noted above, we can think of hawkishness or military assertiveness as reflecting core beliefs about the nature of adversaries in the international system (either benevolent, or malevolent), and hence the efficacy and desirability of military force as an instrument of foreign policy (Glaser, 1992; Herrmann, Tetlock and Visser, 1999; Kahneman and Renshon, 2007; Brutger and Kertzer, 2017). As such, hawks and doves have different reference points: doves view foreign policy through

\textsuperscript{12}The complete regression tables are presented in Appendix §3.3.
Figure 5: Moderating effects of leader characteristics on perceived efficacy of signals

(a) Public threat

(b) Mobilization

Note: Panel a presents the bootstrapped distribution of coefficient estimates from model 7 from Table 2 in Appendix §3.3; panel b presents the bootstrapped distribution of coefficient estimates from model 7 from Table 3 in Appendix §3.3. Each coefficient represents the moderating effect of each orientation and experience on each treatment effect. The plots show relatively strong moderating effects for political orientations like military assertiveness, but generally fail to find evidence in favor of moderating effects of leader-level experiences. The models also control for gender and left-right political ideology, the results for which are omitted here to save space.
the prism of diplomacy, and hawks through the prism of force. Thus, hawks should be especially attuned to military signals that actually involve a display of force, rather than those that merely suggest the possibility of force in the future. By this logic, we should therefore also expect hawks to view military mobilization as more credible than public threats. And indeed, when it comes to judgments of the efficacy of military mobilizations on estimates of resolve in panel (b) on the right, military assertiveness once again has a significant effect, though here in the opposite direction. Where hawkish leaders were relatively unimpressed by signaling based on verbal threats, they took seriously the mobilization of military forces. The findings thus show an asymmetry in how hawks interpret signals in a manner that nicely captures their beliefs in the utility of force: those signals that directly involve military force are seen as more credible than those that engage them only indirectly.

We analyze this pattern further in Table 2, the first two rows of which present a set of difference-in-differences showing the within-subject effect of threatening force compared to the within-subject effect of mobilizing troops, producing different estimates for doves and hawks, respectively, by mean-splitting the military assertiveness scale and estimating separate difference-in-differences within each subgroup. The third row presents the difference-in-difference-in-difference: the difference in estimates of resolve produced by threatening force compared to mobilization, for hawks versus doves. Thus, the first row shows that if you want to signal resolve to a dovish leader, public threats work just as audience cost theorists expect: doves update their assessment of an opponent’s resolve by 8.44 points more when the opponent issues a verbal threat compared to when they mobilize military forces ($p < 0.03$). In fact, supplementary analyses suggest that for the most dovish 22% of our leaders, we cannot reject the null hypothesis that mobilization fails to impact estimates of resolve at all: consistent with the spiral model, doves view military gestures as relatively ambiguous and prone to misperception. In contrast, if you want to signal resolve to a hawkish leader, deeds (mobilization) trump words (public threats): hawks update their assessment of an opponent’s resolve by 3.78 points more in response to mobilization than a public threat ($p < 0.12$). In fact, threats aren’t seen as credible at all by the most hawkish 14% of our leaders. As a result, the difference-in-difference-in-difference (or, the difference in the within-subject effects of public threats compared

---

13 The estimates come from calculating the fitted values from a bivariate regression model regressing the within-subject treatment effect for public threats on a continuous measure of hawkishness, bootstrapping it 1500 times, and identifying the values of militant assertiveness for which the 95% CIs cross zero. When we do the same with the multivariate model shown in model of 7 in Table 3 in Appendix §3.3 (holding all other continuous variables at their means, and all ordinal variables at their medians), we fail to reject the null hypothesis that mobilization fails to affect estimates of resolve for the most dovish 47% of respondents.
to mobilization between doves and hawks) is substantively large (12.22 points) and statistically significant ($p < 0.02$).

<table>
<thead>
<tr>
<th>$\Delta$ in estimate of resolve</th>
<th>Leader type</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>... public threat compared to mobilization</td>
<td>Dove</td>
<td>8.44 ($p &lt; 0.03$)</td>
</tr>
<tr>
<td>... public threat compared to mobilization</td>
<td>Hawk</td>
<td>−3.78 ($p &lt; 0.12$)</td>
</tr>
<tr>
<td>Difference-in-difference-in-difference</td>
<td></td>
<td>−12.22 ($p &lt; 0.02$)</td>
</tr>
</tbody>
</table>

Note: the first two entries are difference-in-differences (the effect of each threat is calculated at the within-subject level, in reference to each participant’s baseline estimate).

The substantive importance of this difference-in-difference-in-difference is worth emphasizing. IR scholars have long been interested in the notion of foreign policy substitutability: the idea that governments have multiple tools they can use to achieve the same objective (Most and Starr, 1984). As two different types of costly signals, public threats and military mobilization should in theory be substitutes for one another: a government eager to credibly signal its resolve in a crisis to an adversary can choose one or the other in order to achieve the same basic goal. And yet, as Milner and Tingley (2015) show in the domestic political context, heterogeneous preferences or beliefs serve as an impediment to substitutability. In the case of costly signaling, different types of signals appear to be interpreted differently by different types of leaders: if only doves find public threats credible, for example, public threats and military mobilization are imperfect substitutes for one another, since different kinds of targets require different types of signals. Signaling correctly thus requires knowing your adversary (Yarhi-Milo, 2014).

Finally, we find mixed evidence for H3b. Leaders who are high on international trust view mobilization of troops and public threats as informative signals of resolve compared to leaders who are low on international trust. However, of these two types of signals, leaders with high level of international trust tend to view mobilization of troops as a more significant indicator of future behavior than public threats, similarly to hawks.

### 4.2.2 Leader experiences

Thus far we have shown that that leaders’ orientations — in particular, their beliefs about the desirability and efficacy of force, as well as levels of international trust — affect how they interpret costly signals. In contrast, the results in the top panel of Figure 5 suggest little support for the notion that the judgment of our leaders was affected by their military and political experiences. Consider
military or combat experience as an example. A number of political scientists have pointed to military experience as an important predictor of leaders’ conflict behavior, suggesting that time served in the military, experiences in combat, or military service without the experience of combat, shape how leaders evaluate costs and benefits regarding decisions about military force (e.g., Sechser, 2004; Horowitz, Stam and Ellis, 2015). Due to conscription, 95% of our participants served in the military, so in lieu of focusing on military service, we turn to the presence or absence of active combat experience. Importantly, we find that neither combat experience, nor military experience without combat, significantly affects how our leaders interpret signals. Across both types of costly signals, the only leader-level experiential variable that approaches statistical significance concerns age or time in office, with older leaders, or leaders who served a greater number of terms in office, perceiving military mobilization as a less credible signal. Indeed, a set of Wald tests in Appendix §3.3 generally fails to find evidence of a significant reduction in model fit if we drop these experiential variables altogether. Our results do not suggest that experiences are immaterial — the divergent magnitudes of treatment effects between the leader and public surveys show that experience matters, even if the results differ in size rather than sign — but rather, that they cannot explain the variation we detect within our leader sample.

Why might we fail to find evidence in favor of leader experiences here? Two points are worth noting. First, one interpretation of our results might attribute the relatively weak effects of leader experiences to post-treatment bias (King and Zeng, 2007): if we expect that leaders’ experiences shape their general orientations about beliefs about international politics, which then act as prisms through which they process new information and form policy views, estimating the effects of experiential variables while also controlling for foreign policy orientations can erroneously bias our estimate of the latter. And yet, when we reestimate our models while dropping the effects of orientations in supplementary analyses in Tables 4-5 of Appendix §3.3, leader experiences still fail to reach conventional thresholds of statistical significance.

Second, we might fail to find evidence in favor of leader experiences because of measurement error: perhaps the measures of leader experiences we use here are poor proxies for the underlying construct. And yet, the measures of leader experience we use here include many of the same variables used elsewhere in the study of leaders. Our findings thus encourage IR scholars to focus on the scope conditions for theories of leader experiences by exploring how past military or political experiences shape political judgments, strategic assessments, and behavior.
5 Conclusion

In the above discussion, we offered unique experimental evidence from elite foreign policy decision-makers to answer three questions. First, do costly signals work? Second, are some types of costly signals more effective than others in sending credible signals of resolve? And finally, do leader-level characteristics matter in explaining variation in perceptions of the signals’ efficacy?

Our main results are thus threefold. First, costly signaling works: although military mobilization has been criticized for being ineffectual, and public threats criticized for being empty, in our experiments, both types of signals appear to work in the manner that our theoretical models predict. Second, utilizing the “between-subjects” part of our design, we fail to find evidence that one type of signal is more effective than the other. Nonetheless, in suggesting that military mobilization is just as effective as public threats, our results differ from recent experimental evidence that suggests that sunk costs are not perceived as effective by observers (Quek, 2016). One potential explanation for the divergent conclusions may be different experimental designs, since we randomly assign whether military mobilization takes place in order to identify its causal effect. Our findings thus suggest we shouldn’t abandon hope in sinking costs just yet.

Of course, caution should be exercised in over-learning from any set of studies. While we used a relatively large sample of leaders for this type of research, and were able to replicate our main findings on a much larger, representative sample of the Israeli public, there is still much for future work to fill in. For example, other studies can help by varying attributes of the design in meaningful ways, for example by including conditions with cheap talk signals; our results show that leaders update in response to costly signals, but it is also possible that leaders update in response to costless ones as well. Other variations on our design might help to contextualize our results by exploring how much the stylized nature of our experiment accounts for the size of our effects. One possibility is that it exaggerates them relative to what they would be in a natural setting by focusing the attention of subjects on only a few pieces of information, while an alternative is that elements in the “real world” (e.g., the dosage of the treatments, stakes, etc.) might actually contribute to larger substantive effects. The size of our effects may also be affected by which country our mass public participants have in mind when they think of a “democracy,” which in turn has implications for the generalizability of our results; perhaps the absence of a democratic advantage in signaling is a result of our participants calling to mind a particularly weak country when presented with the democratic
In demonstrating the efficacy of public threats, our results also vitiate one of the main tenets of audience cost theory, which has been subject to considerable skepticism. Interestingly, although the canonical audience cost model consists of a contest between two states, existing experimental research on audience costs has only focused on the judgments of a single state’s domestic audience (e.g., Tomz, 2007; Trager and Vavreck, 2011; Kertzer and Brutger, 2016). Our experiment thus represents the first study of which we are aware to explore audience cost theory through the eyes of foreign observers. In so doing, our findings suggest that some critiques of audience costs may be misplaced: as long as foreign audiences believe that public threats are costly, they can be effective even if domestic audiences would not actually punish backing down, and even if leaders can shape the reactions of their constituents (Paper #2 This Issue, Paper #3 This Issue, Paper #5 This Issue). However, although our findings about the credibility of public threats is consistent with audience cost theory, in our public study we fail to find evidence that democracies are uniquely advantaged in terms of their ability to use public threats. In this sense, our results are consistent with claims that democratic signaling advantages may be overstated (e.g., Weeks, 2008; Downes and Sechser, 2012).

Third, although theories of costly signals have tended to view credibility as purely situational and based on actors’ payoff structures, as Mercer (2010) shows, credibility is a belief, varying across observers. We find the main predictors of this variation are leaders’ orientations: public threats are less effective against hawks, and individuals high in international trust see public threats as more credible than individuals low in international trust. Meanwhile, hawks are inclined to see “deeds” as particularly effective signals, viewing military mobilization as a more credible signal of resolve than do doves. These divergent reactions means that different types of leaders will perceive the same signal in multiple ways; costliness is in the eye of the beholder, such that costly signals are not necessarily substitutable for one another. The fact that the dispositions we focus on here have also been found to exert important effects in mass public opinion about foreign affairs (e.g., Kertzer and Brutger, 2016) point to the extent to which the orientations found to structure mass political attitudes also play important roles in elite political judgment as well.

In contrast, we generally fail to find evidence that leaders’ military and political experience shapes how they calculate credibility, and show that these null results are not due to post-treatment bias. Our findings thus remind us that if leader-level experiences matter because they shape leaders’ preferences and beliefs, the processes through which beliefs come to be formed are sufficiently com-
plex that it remains worthwhile for IR scholars to study them in their own right as well (Saunders, 2011). Our results thus accord with an older wave of research on leaders focusing directly on mental models (e.g., George, 1969) rather than the formative events believed to be responsible for them.

In this sense, our findings raise some provocative questions for our theories of IR more generally. Unlike psychological work in IR which has long been interested in leaders’ perceptions, the rationalist IR literature traditionally shied away from placing a heavy emphasis on leaders, because in order for leaders to matter, they can’t all be the same (Jervis, 2013). Not only does individual heterogeneity make our models less analytically tractable, but it also poses a dilemma for an intellectual tradition that perceives individual-level variation as a kind of noise or error: from a rationalist perspective, why should two actors in an identical strategic situation behave differently from one another? (Lake and Powell, 1999; Kertzer, 2016; Rathbun, Kertzer and Paradis, 2017).14 Perhaps one reason why the recent wave of leader-level literature has tended to embrace experiential variables, then, is because attributing divergent leader behavior to divergent prior experiences “rationalizes” heterogeneity, rendering it consistent with a clean rationalist story in which actors who had similar prior experiences should respond to a given strategic environment in much the same way. Yet, our results suggest limitations to this story in that, even if experiences shape orientations, there remains important residual variation that the types of experiences we often study in IR cannot yet explain.

Thus, we end by considering where these orientations might originate. The most obvious possibility is that hawkishness is still shaped by prior experiences, but in complicated, nonlinear ways that are too subtle to be picked up by our research design. It’s also possible that IR scholars could still be correct in asserting that orientations are shaped by experiences, but those experiences may differ from those that we typically examine. Instead of time in office or combat or rebel experience, it may well be “formative” experiences (à la George and George, 1964) that matter most. More broadly, political orientations may simply be extensions of attributes that are pre-political but spill over into the political domain, such as values (Rathbun et al., 2016) or beliefs about the way the world works (Renshon, 2008). Finally, orientations might be shaped by genetic factors (McDermott, 2014). These categories are neither exclusive nor exhaustive, but they do provide a starting place for what is sure to be a burgeoning area of research directed at disentangling the role of experiences and orientations in foreign policy behavior.

14 Notice, for example, how selectorate theory (and what the introductory paper calls the “institutional leadership school” more generally) has gotten around this dilemma by suggesting that leaders’ domestic political situation varies; it is not that leaders themselves are different, but rather that they face different incentive structures.
References


Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton, NJ: Princeton University Press.
Tying hands, sinking costs, and leader attributes:
Supplementary appendix

August 4, 2017

Contents

1 Recruitment protocol 2
  1.1 Recruitment letter to Knesset Members ........................................ 2
    Figure 1: Recruitment Letter ........................................ 2
  1.2 Recruitment procedures ...................................................... 2
  1.3 Participant verification protocol ........................................ 4

2 Study instrumentation 6
  2.1 Experimental protocol (translated to English) ................................ 6
  2.2 Individual difference measures ........................................... 7
    2.2.1 Military assertiveness ........................................... 7
    2.2.2 International trust ............................................. 7
    2.2.3 Right-wing ideology ............................................ 7

3 Supplementary analysis 8
  3.1 Representativeness and survey non-response ............................ 8
    Table 1: Gauging the representativeness of the sample ................... 9
  3.2 Generalizability outside the Israeli context ............................ 10
  3.3 Leader characteristics and robustness checks .......................... 13
    Table 2: Effect of leader characteristics on perceived efficacy of public threats ..... 14
    Table 3: Effect of leader characteristics on perceived efficacy of military mobilization 15
    Table 4: No evidence for post-treatment bias for threats ................. 16
    Table 5: No evidence for post-treatment bias for mobilization .......... 16

4 Israeli public sample 18
  Table 6: Israeli public sample demographics ................................ 18
  4.1 Elite-public differences in political orientations ....................... 18
    Table 7: Sample comparison ........................................... 19
1 Recruitment protocol

1.1 Recruitment letter to Knesset Members

Figure 1: Recruitment Letter

1.2 Recruitment procedures

We began the recruitment process by compiling a dataset of all 415 individuals who had served as members of the Parliament of Israel (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 in Appendix §1.1), 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.

1.3 Participant verification protocol

We took several steps to increase our confidence that the current and former decision-makers participated 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 members of his or her staff. We explained that the code we provided to access the on-line survey was 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.
2 Study instrumentation

2.1 Experimental protocol (translated to English)

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%?

2.2 Individual difference measures

2.2.1 Military assertiveness

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

2. The use of force generally makes problems worse

3. Rather than simply reacting to our enemies, it’s better for us to strike first.

All scaled from 1 (strongly agree) to 5 (strongly disagree). Items 1 and 3 reverse-coded such that higher scores indicated higher level of military assertiveness.

2.2.2 International trust

1. Some people say that Israel can trust other nations, while others think that Israel can’t be too careful in dealing with other nations. Where would you place yourself on this scale from 1 (Israel can count on other countries) to 7 (Israel cannot count on other countries)?

Reverse-scored such that higher scores indicate more higher levels of trust.

2.2.3 Right-wing ideology

1. There is much talk of “left” and “right” in politics. How would you rate yourself on a left-right scale, from 1 (right) to 7 (left)?

Reverse-scored such that higher scores indicate more right-wing ideology.
3 Supplementary analysis

3.1 Representativeness and survey non-response

There are two ways of thinking about the representativeness of our elite sample. The first asks how our participants compare to the complete population of individuals who served in the Knesset from 1996 to the present. The second asks how 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 1 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.

Israeli legislative politics features a number of characteristics that makes calculating an summary partisan representativeness score somewhat complex, including a high degree of fragmentation, parties frequently splintering and forming, frequent party switching, and the presence of Arab and religious parties that cannot be cleanly positioned on unidimensional partisan space. As a result, after coding all parties in the Knesset from 1996-2015 as being either left, center, right, Arab, religious, or none of the above, we look at partisan representativeness in two different ways. First, we focus on the representation of MKs from Arab parties (e.g. Hadash, Ra’am, Ta’al, Balad, the Joint Arab list, etc.), and religious parties (e.g. Shas, United Torah, Tkuma, Jewish Home, etc.), which are orthogonal to the left-right spectrum and thus are separated from the main analyses. We find that although both Arab and religious MKs are included in the sample, they are both slightly underrepresented: MKs from Arab parties make up approximately 7% of the population of MKs in this time range, but only 3% of the sample; MKs from religious parties make up approximately 19% of the population of MKs in this time frame, but only 7% of the sample. Second, we take the
members who were exclusively associated with parties on the main ideological spectrum (left, center, and right), and calculate a left-right ideology score averaging across their terms in office within the sampling period, such that the lowest score (1) is assigned to an MK who consistently represented a left-wing party, and the highest score (3) is assigned to an MK who consistently represented a right-wing one. So, for example, an MK that served two terms in a left-wing party and a third term in a centrist party would be coded as being slightly more conservative (1.33) than an MK that served all of their terms in a left-wing party (1), and much less conservative than an MK who served their terms in right-wing parties (3). Models 2 and 4 in Table 1 show that our sample is slightly more left-wing than the population of MKs in this time period as a whole, but interestingly, that this skew is not due to non-response bias, in that left-leaning MKs were not significantly more likely to participate in the survey than right-leaning ones. The partisan difference thus appears to stem from the probability of entering the sampling frame rather than the probability of response.3

religious classification for the purpose of this analysis, and would be excluded from the left-right ideology analysis discussed below.

3Supplementary analyses suggest that the average left-right ideology amongst deceased MKs is more conservative (2.4) than living MKs (2.0), potentially explaining some of this effect.

Table 1: Gauging the representativeness of the sample

<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>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
3.2 Generalizability outside the Israeli context

As we note in the main text, our choice to study how leaders understand costly signals using Israeli leaders is valuable for a number of reasons, both theoretically and methodologically. First, because of the Israeli parliamentary system, the majority of the executive branch are also elected members of the Knesset, such that the Knesset is comprised of policy makers who are directly involved in decisions about the use of force. Second, because of Israel’s relatively short election cycles, it is common for former members of the executive to later become members of the opposition in the Knesset, such that nearly all of Israel’s current Executive branch members were at one point members of the opposition. Thus, even those participants in our study who are currently opposition members have either been members of the Executive in the past, or are likely to be so in the future, such that we are effectively sampling current, former, and potentially future members of the Executive. Third, if the goal of our project is to examine how decision-makers perceive costly signals in foreign policy crises, it makes the most sense to do so in a country where these questions about war and peace are salient and realistic, where decision-makers have had to wrestle with exactly these kinds of issues — this, after all, is what is unique about studying elite decision-makers rather than college sophomores. Israel is a country where these “use of force” decisions are extremely common, rendering these questions particularly salient. Indeed, during the time frame in which our participants were in office (from 1996 onwards), Israel was involved in 16 Militarized Interstate Disputes (MIDs – Jones, Bremer and Singer, 1996); many our participants were involved in the decision-making process for many of these incidents. And, since we can expect that earlier experiences of Israeli conflict might have shaped leaders’ beliefs, we note that Israel has been involved in 128 MIDs since the state’s founding, and seven militarized compellent threats (Sechser, 2011).

Nonetheless, the value of the Israeli case does not preclude the need for thoughtful reflection concerning the generalizability of our findings. Typically this question is understood as implicating the tradeoff between internal and external validity. As Imbens (2010, 403) notes, the statistical literature has generally “emphasized internal validity over external validity, with the view that a credible estimate of the average effect for a subpopulation is preferred to an estimate of the average for the target population with little credibility.” In political science, McDermott (2011) takes the same view, noting that it one needs to know whether an effect exists at all before one can meaningfully ask whether it travels to other contexts. In a sense, though, the question is more subtle than how the tradeoff between internal and external validity is often understood by experimental critics: in their important critique of whether survey experiments lack external validity, for example, Barabas and Jerit (2010) show that one weakness of survey experiments is that they exaggerate treatment effects, because participants outside of the survey context are often less likely to receive the treatment in
the first place. In our case, however, the unique salience of foreign policy crises in Israel actually *mitigates* this aspect of external validity concerns!

Ultimately, the best means of exploring how our results might generalize outside the Israeli context is to address the question empirically, fielding the study in other democratic and nondemocratic regimes – a task we encourage interested readers to pursue. Short of that, however, we can address the question theoretically. We thus think through how our results might generalize outside the Israeli context by following best practices suggested by many scholars of comparative politics and qualitative work in IR (George and Bennett, 2005) by focusing on *within-case* variation: although Israel may differ from other countries in important ways, we can exploit within-country variation along the dimensions in which Israel differs to provide some perspective on how one might expect results to differ in other contexts. In particular, there are at least three features of the Israeli context that could potentially affect our ability to draw inferences about leaders in other countries. The first concerns the widespread prevalence of military service in Israel: unlike in many other countries, all Jewish Israeli citizens are required by law to serve in the IDF, many of whom experience combat firsthand. As noted in the main paper, although compulsory service in the IDF means we do not have sufficient variation in our sample of decision-makers to estimate heterogeneous treatment effects with respect to military service, there is considerable variation in combat experience. In the main paper, we exploit this variation to estimate whether there are heterogeneous treatment effects with respect to combat experience, which also serves as a proxy for whether our results would hold in countries where leaders might have less military experience. Yet as the main results show, participants’ combat experience does not appear to affect how they update based on different types of costly signals.

The second concerns the relatively hardline nature of Israeli politics on security issues, which may make our participants consistently more hawkish than decision-makers in other countries. Here, we do find significant evidence of treatment heterogeneity in our sample, with hawks more sensitive to sunk cost signals of military mobilization, but less persuaded by verbal threats than doves. However, if the expectation is that Israeli leaders are unusually high in military assertiveness compared to leaders in other countries, this should make it harder to find an effect in Israel than in countries where a wider range of attitudes towards military force are represented.

The third concerns the high salience of foreign policy and security issues in Israeli politics, which may make our decision-makers more experienced and knowledgeable about assessing the credibility of threats and decisions about the use of force. As noted in the main text, this actually increases the validity of our participants’ responses, since we are asking them to engage in the very behavior in which they are more likely to have expertise. Nonetheless, in the main paper we exploit variation
in foreign affairs experience among our participants to estimate heterogeneous treatment effects. Interestingly, at least when we measure foreign affairs experience formally (that is, in terms of serving on the Foreign Affairs and Defense Committee) we find no evidence that more experienced participants interpret signals differently than less experienced ones, although it is possible that in countries where foreign policy is less salient, formal foreign affairs experience might matter more. Thus, although future scholarship should extend our findings by carrying out similar studies in other countries, when we focus on the three dimensions above, we fail to find evidence of heterogeneous treatment effects.
3.3 Leader characteristics and robustness checks

Tables 2-5 present a series of supplementary analyses for the leader-level analyses from the main text. Due to the small sample size, and intercorrelation between a number of our measures of experience, models 1-6 of Table 2 present a series of regression models estimating the effect of each leader experience variable separately on the perceived efficacy of public threats; the complete model is presented in model 7. Table 3 does the same for the effects of leader experiences on the perceived efficacy of military mobilization. In both cases, the leader experience variables appear to be relatively unimportant in explaining variation in how leaders calculate credibility, while the effect of leader orientations are substantively stronger, particularly military assertiveness and international trust in Table 2. Indeed, a pair of Wald tests fail to find evidence of a significant reduction in model fit if the experiential variables are dropped altogether, apart from age in the military mobilization treatment, which implicates a variety of non-experiential mechanisms as well.\footnote{For public threats, a Wald test comparing the unrestricted model from model 7 of Table 2 with a restricted model without the experiential variables (combat experience, terms in office, foreign affairs experience, elite experience, and age) fails to find evidence of a significant reduction in model fit \((F = 0.85, p < 0.52)\); an equivalent test for mobilization using model 7 of Table 3 suggests a significant reduction in model fit \((F = 2.29, p < 0.07)\), but this is due to the significant effect of age, rather than any of the other experiential variables, and if we replicate the analysis while also retaining age, the reduction in fit is no longer significant \((F = 1.03, p < 0.41)\).}

We also carry out several additional robustness checks. First, to show that the null effects of leader experiences aren’t simply due to post-treatment bias induced by estimating the effects of experiences and orientations simultaneously, Tables 4-5 re-estimate the models from Tables 2-3, this time omitting the leader orientation variables. Importantly, the leader-level variables continue to lack statistical significance; a Wald test comparing model 7 from Table 3 with model 7 from Table 5 confirms dropping orientations from the model produces a significant reduction in fit \((F = 4.917, p < 0.007)\), while a Wald test comparing model 7 from Table 2 with model 7 from Table 4 fails to find evidence of a significant reduction in fit \((F = 2.026, p < 0.130)\), although the sample size is small in both cases.

Second, we seek to adjudicate between two different explanations for the divergent results we see between hawks and doves in the results. The account offered in the main text holds that hawks and doves differ from one another in how they interpret the signals. An alternative account would be that hawkish and dovish leaders interpret signals similarly, but have differing baseline expectations of how likely other actors are to stand firm in crises - those who update more may simply do so because they have more diffuse priors. We find this explanation less plausible, since it cannot explain why each type of costly signal has different effects, but the within-subject component of the experimental design lets us test this account directly. We calculate the strength of each participant’s priors from by calculating the distance of their baseline resolve estimate from 50%. Importantly, the strength
Table 2: Effect of leader characteristics on perceived efficacy of public threats

<table>
<thead>
<tr>
<th>Experience</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combat experience</td>
<td>-6.663</td>
<td></td>
<td></td>
<td></td>
<td>-4.966</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.586)</td>
<td></td>
<td></td>
<td></td>
<td>(5.246)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Military experience, no combat</td>
<td>6.519</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.679)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terms</td>
<td>0.465</td>
<td></td>
<td></td>
<td></td>
<td>-0.246</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.140)</td>
<td></td>
<td></td>
<td></td>
<td>(1.557)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreign affairs experience</td>
<td></td>
<td>-4.255</td>
<td></td>
<td></td>
<td>-5.676</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.700)</td>
<td></td>
<td></td>
<td>(6.289)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elite experience</td>
<td>6.992</td>
<td></td>
<td></td>
<td></td>
<td>10.375</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.855)</td>
<td></td>
<td></td>
<td></td>
<td>(9.469)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.221</td>
<td>-0.135</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.217)</td>
<td>(0.259)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Orientations                    |         |         |         |         |         |         |         |
| Military assertiveness          | -36.061*| -37.782**| -38.009*| -34.410*| -38.687**| -38.433**| -38.031*|
| Ideology                        | 7.522   | 7.185   | 6.975   | 10.462  | 4.548   | 10.559  | 10.450  |
| N                               | 42      | 42      | 43      | 43      | 43      | 43      | 42      |
| R²                              | 0.212   | 0.208   | 0.169   | 0.178   | 0.188   | 0.188   | 0.263   |

*p < .1; **p < .05; ***p < .01
Table 3: Effect of leader characteristics on perceived efficacy of military mobilization

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experience</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combat experience</td>
<td>-1.972</td>
<td></td>
<td></td>
<td></td>
<td>-1.136</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.696)</td>
<td></td>
<td></td>
<td></td>
<td>(4.501)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Military experience, no combat</td>
<td>5.973</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.167)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terms in office</td>
<td></td>
<td>-1.910*</td>
<td></td>
<td></td>
<td>-2.730*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.987)</td>
<td></td>
<td></td>
<td>(1.425)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreign affairs experience</td>
<td>-0.408</td>
<td></td>
<td></td>
<td></td>
<td>4.463</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4.369)</td>
<td></td>
<td></td>
<td>(4.531)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elite experience</td>
<td></td>
<td>-4.341</td>
<td></td>
<td></td>
<td>5.399</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.038)</td>
<td></td>
<td></td>
<td>(7.548)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td>-0.557***</td>
<td></td>
<td></td>
<td>-0.372</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.201)</td>
<td></td>
<td></td>
<td>(0.227)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Orientations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Military assertiveness</td>
<td>-40.243**</td>
<td>40.280**</td>
<td>41.022**</td>
<td>38.120**</td>
<td>36.766**</td>
<td>52.176***</td>
<td>52.519***</td>
</tr>
<tr>
<td>Male</td>
<td>-2.889</td>
<td>-0.572</td>
<td>-3.466</td>
<td>-3.055</td>
<td>-2.643</td>
<td>-2.940</td>
<td>-4.083</td>
</tr>
<tr>
<td></td>
<td>(5.848)</td>
<td>(6.094)</td>
<td>(5.305)</td>
<td>(5.563)</td>
<td>(5.555)</td>
<td>(5.061)</td>
<td>(5.510)</td>
</tr>
<tr>
<td>N</td>
<td>41</td>
<td>41</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>41</td>
</tr>
<tr>
<td>R²</td>
<td>0.218</td>
<td>0.243</td>
<td>0.263</td>
<td>0.188</td>
<td>0.199</td>
<td>0.328</td>
<td>0.407</td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01
Table 4: Robustness check: no evidence of post-treatment bias for threats

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combat experience</td>
<td>-5.963</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-6.233</td>
</tr>
<tr>
<td></td>
<td>(4.687)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(5.348)</td>
</tr>
<tr>
<td>Military experience, no combat</td>
<td>4.931</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.785)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terms in office</td>
<td>0.706</td>
<td></td>
<td>0.078</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.959)</td>
<td></td>
<td>(1.454)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreign affairs experience</td>
<td>-3.707</td>
<td></td>
<td>-8.236</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.851)</td>
<td></td>
<td>(5.928)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elite experience</td>
<td>5.649</td>
<td>8.972</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.160)</td>
<td>(9.227)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.049</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.199)</td>
<td>(0.258)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>7.982</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.844)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>12.000***</td>
<td>6.536**</td>
<td>5.891</td>
<td>10.769**</td>
<td>6.439**</td>
<td>11.107</td>
<td>7.237</td>
</tr>
<tr>
<td></td>
<td>(3.714)</td>
<td>(2.826)</td>
<td>(3.759)</td>
<td>(4.091)</td>
<td>(2.867)</td>
<td>(12.279)</td>
<td>(13.253)</td>
</tr>
<tr>
<td>N</td>
<td>43</td>
<td>43</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>43</td>
</tr>
<tr>
<td>R²</td>
<td>0.038</td>
<td>0.025</td>
<td>0.012</td>
<td>0.013</td>
<td>0.019</td>
<td>0.001</td>
<td>0.124</td>
</tr>
</tbody>
</table>

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

Table 5: Robustness check: no evidence of post-treatment bias for mobilization

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combat experience</td>
<td>0.606</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.602)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(5.072)</td>
<td></td>
</tr>
<tr>
<td>Military experience, no combat</td>
<td>3.191</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.900)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terms in office</td>
<td>-1.376</td>
<td></td>
<td>-1.724</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.042)</td>
<td></td>
<td>(1.603)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreign affairs experience</td>
<td>2.366</td>
<td></td>
<td>7.646</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.261)</td>
<td></td>
<td>(5.118)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elite experience</td>
<td>-4.784</td>
<td>-1.408</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.913)</td>
<td>(8.260)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.305</td>
<td>-0.242</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.200)</td>
<td>(0.254)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>-0.192</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>6.357*</td>
<td>5.900**</td>
<td>10.697***</td>
<td>5.313</td>
<td>7.960***</td>
<td>25.714**</td>
<td>20.976</td>
</tr>
<tr>
<td>N</td>
<td>41</td>
<td>41</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>41</td>
</tr>
<tr>
<td>R²</td>
<td>0.0004</td>
<td>0.011</td>
<td>0.040</td>
<td>0.007</td>
<td>0.015</td>
<td>0.053</td>
<td>0.125</td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01
of participants’ priors and their levels of hawkishness are independent of one another \((r = -0.025)\); if we define the threshold between hawks and doves by mean-splitting hawkishness, hawks have an average baseline level of certainty of 17.38, and doves of 17.48. Alternately, rather than focus on the strength or diffusion of participants’ priors more generally, we can simply look at the average baseline estimate offered by hawks and doves (e.g. perhaps hawks assume foes are more resolute than doves do?). Here too, though the correlation between participants’ baseline assessments and their level of hawkishness is relatively weak \((r = 0.07)\); mean-splitting hawkishness, it appears that hawks offer slightly higher estimates than doves (56.4 vs 51.9), but the difference is not significant: in a bivariate regression model of baseline estimates on hawkishness, the associated p-value for the hawkishness coefficient is \(p < 0.51\). These tests thus offer further evidence that the results for hawkishness are not driven by differing priors, but rather, differing interpretations of the signals themselves.

Finally, given the effects of analogical reasoning (Khong, 1992) and the outsized role that a series of formative historical events — e.g. the six-day war in 1967, the Yom Kippur war in 1973 — play in Israeli political culture, we tested for generational effects (Kertzer, 1983). Were individuals old enough to remember the 1973 war, for example — a conflict in which Israel was surprised by military mobilization — more likely to take military mobilization seriously? Our results suggest the opposite, in that younger MKs update more in response to military mobilization than older ones. Given, however, the relatively small sample size (relatively few respondents in our sample were born after 1973, and because of longitudinal changes in the composition of the Knesset over time, younger MKs differ from older ones in a number of ways), we lack the data to test this hypothesis conclusively.
4 Israeli public sample

Table 6: Israeli Public Sample

<table>
<thead>
<tr>
<th>Public sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
</tr>
</tbody>
</table>

**Education:**
- No High School degree | 2%
- High School degree | 33%
- Some college | 23%
- College degree | 27%
- Masters degree | 14%
- Doctoral degree | 2%

**Military Experience:**
- Did not serve | 22%
- Served, no active combat | 50%
- Combat experience | 28%

**Location:**
- Jerusalem | 10%
- Tel Aviv | 12%
- Central Zone | 18%
- Haifa | 15%
- Northern Region | 12%
- Southern Region | 12%
- Lowland | 10%
- Sharon Area | 8%
- Yehuda and Shomron | 3%

**Religiosity:**
- Secular | 60%
- Traditional | 19%
- Religious | 15%
- Orthodox | 5%

**Birth Country:**
- Israel | 81%
- Former USSR | 10%
- Other | 9%

<table>
<thead>
<tr>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Military Assertiveness</td>
</tr>
<tr>
<td>Right Wing Ideology</td>
</tr>
<tr>
<td>Hawkishness (Arab-Israeli conflict)</td>
</tr>
<tr>
<td>International Trust</td>
</tr>
</tbody>
</table>

4.1 Elite-public differences in political orientations

In addition to standard demographic data, subjects completed questionnaires relating to military assertiveness (described in the main text), political ideology, hawkishness (with respect to the Arab-
Israeli conflict) and international trust. Our political ideology item asks subjects to classify themselves along a single dimensions from “left” to “right” in politics, while the “Hawkishness” measure asks subjects to do the same specifically with respect to the Israeli-Arab conflict. As before, our international trust measure is adapted from the work of Brewer et al. (2004), which finds that generalized trust of other countries in the international system helps to structure beliefs about the foreign policy arena.

Table 7 compares our samples along several dimensions, all scaled from 0 (min response) to 1 (max response). First, we note that — within the leader sample — current and former Knesset members do not differ very much on measured covariates. In fact, the only statistically significant difference between the two subsamples of leaders is that current leaders are younger (by about 12 years) compared to former leaders. The differences along the ideational dimensions are all rather small, and none of them are statistically significant.

Perhaps unsurprisingly, our sample of leaders is significantly older (by about 20 years) than respondents from the Israeli public. Some of this is driven by the former Knesset members, who outnumber the younger, current members of Parliament in our sample. In fact, current Knesset members in our sample average 52 years (64 years for former members), far closer to the average age of the Israeli public sample (42).

Our Israeli leaders are notably less conservative, less hawkish with respect to the Arab-Israeli conflict and more trusting of the international community than the general public. With respect to left-right political ideology, Knesset members averaged a score of 0.45, a score that places them nearly an entire standard deviation less conservative than the mean score in the Israeli public sample. With respect to the Arab-Israeli conflict, our leaders were an entire standard deviation below the mean for the public sample.

<table>
<thead>
<tr>
<th></th>
<th>Knesset</th>
<th>Public</th>
<th>Elite–Public Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Current</td>
<td>Former</td>
<td>Overall</td>
</tr>
<tr>
<td>Age</td>
<td>52.2</td>
<td>64.4</td>
<td>61.4</td>
</tr>
<tr>
<td>Military Assertiveness</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>Right Wing Ideology</td>
<td>0.47</td>
<td>0.44</td>
<td>0.45</td>
</tr>
<tr>
<td>Hawkishness (Arab/Israeli)</td>
<td>0.44</td>
<td>0.37</td>
<td>0.39</td>
</tr>
<tr>
<td>International Trust</td>
<td>0.37</td>
<td>0.41</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 7: Sample Comparison: Statistically significant differences in means between public sample and leader (overall) sample depicted in bold. p-values calculated via Wilcoxon rank-sum tests.
References


