Tying Hands, Sinking Costs, and Leader Attributes

Keren Yarhi-Milo¹, Joshua D. Kertzer², and Jonathan Renshon³

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 international relations 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.

Keywords
leaders, signaling, political psychology, experiments

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 international relations (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 carry out them (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 a 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 criticism 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 article, 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 histories 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, 2018; Horowitz et al. 2018) 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 is just as persuasive to receivers as 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 Sechser 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 toward 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

Table 1. Hypotheses.

| Hypothesis 1 | 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. |
| Hypothesis 2A | 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. |
| Hypothesis 2B | Leaders will view the adversary’s public threat to escalate the conflict as an equally or less credible signal of resolve compared to the adversary’s mobilization of troops. |
| Hypothesis 3A (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. |
| Hypothesis 3B (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. |
| Hypothesis 4A (Experience: Bayesian updaters) | Leaders with more experience should view costly signals as more informative signals of resolve than leaders with less experience. |
| Hypothesis 4B (Experience: biased learners) | Leaders with more experience should be less likely to update from costly signals of resolve than leaders with less experience. |
matter—IR scholars shouldn’t restrict the study of these characteristics to experimental 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.

**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 such as 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, 148-49). 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 (see also Schelling [1960, 1966]). The ability of 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 between resolute and irresolute states.

Both of these models of signaling have become enormously influential in IR. Scholars have applied sunk-cost models to everything from the study of covert interventions (Carson and Yarhi-Milo 2017) 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 offers 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 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.

There are at least three clusters of reasons why the empirical record for these costly signaling mechanisms is 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 data sets IR scholars typically use may be inappropriate for questions relating to signaling. Downes and Sechser (2012) show that the Militarized Interstate Disputes (MID) and International Crisis Behavior data sets 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): “[if] leaders care about reputation (or domestic support), we can expected the observed effects [of backing down] to be biased towards zero.”

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 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 harkens 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) 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 IR. 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.

Hypotheses

Of the three broad questions that are the focus of this article, the first presents a relatively stark, clear hypothesis (see Table 1 for the complete list of hypotheses). 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 (Hypothesis 1). 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 (Hypothesis 2) 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 (for 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 Hypotheses 1 and 2. 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 (Hypothesis 3). Although there are a wide array
of orientations one could choose from, for reasons of tractability, we focus on two here. The 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 2018). The first concerns their beliefs about the nature of adversaries in the international system, with hawks subscribing to a “deterrence mind-set” 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 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 see military gestures as inherently ambiguous and thus perceive them as 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” (Hypothesis 3a).

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 a 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 toward 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 by 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 (Hypothesis 3b).

Whereas dispositional orientations offer one important lens through which leaders process information, another optic frequently studied in the IR literature (and in the other articles in this issue: Horowitz et al. 2018) 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 United States 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 1991 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 affects not only 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 (Hypothesis 4a). 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 preexisting 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 than inexperienced leaders; they might even exhibit more resistance to updating their beliefs in response to costly signals than inexperienced leaders (Hypothesis 4b).

**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 was in office (from 1996 onwards), Israel was involved in 16 MIDs—and we can reasonably expect that our subjects were 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, for example, the US 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 follow-up study on a representative sample of the Jewish public in Israel.3 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 on 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 representative democracies. Additionally, since all Israeli citizens are required by law to enlist in the military at the age of eighteen,4 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 Online Appendix §3.2.

**Experimental Samples**

The Knesset subjects. We fielded the study from July to October 2015. Of 288 potential subjects, 89 participated, leaving us with a 31 percent response rate, relatively high for this type of research (e.g., Hafner-Burton, LeVeck, and Victor 2015; Bayram 2017). The Knesset sample is described in Table 2, and our recruitment procedures are described in Online Appendix §1. In total, 25 percent of our
participants were currently serving in the Knesset; the rest of the sample (75 percent) was composed of former Knesset members. They were also highly experienced in IR-relevant contexts: 64 percent had active combat experience, and 67 percent 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 three terms in parliament (with some as high as nine terms). While not all our sample had ministerial experience, 42 percent of our sample had served at least as a deputy minister, and 12 percent 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 complicated 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 onward 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 Online Appendix §3.1, we conduct supplementary analyses showing that our sample is fairly representative, although

Table 2. Knesset Sample.

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

<table>
<thead>
<tr>
<th>Gray</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>61</td>
<td>10.7</td>
</tr>
<tr>
<td>Terms in Knesset</td>
<td>3.0</td>
<td>2.1</td>
</tr>
<tr>
<td>Military assertiveness</td>
<td>0.61</td>
<td>0.20</td>
</tr>
<tr>
<td>Right-wing ideology</td>
<td>0.45</td>
<td>0.24</td>
</tr>
<tr>
<td>Hawkishness (Arab–Israeli conflict)</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>International trust</td>
<td>0.40</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Note: \(N = 89\). Individual differences in bottom four rows scaled from 0 to 1.
not surprisingly, current Knesset members were less 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 US 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.

The Israeli public. Our follow-up 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 Online Appendix §4.

The Experiment

Our experiment—the structure of which is depicted in Figure 1—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 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 Online 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 subjects 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, an issue central to the study of signaling. After the experimental components of the study, subjects completed a demographic and dispositional questionnaire (reproduced in Online Appendix §2.2).

**Leaders and Costly Signals**

*Do Leaders Update Beliefs about Resolve in Response to Signals?*

Our first task is to evaluate Hypothesis 1, 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 Hypothesis 1 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 2 plots the average treatment effects for the leaders in the Knesset sample. We find strong support for the notion that

**Figure 2.** Effect of costly signals. 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.
leaders update their estimates of opponent’s resolve when costly signals are sent (Hypothesis 1). Leaders saw public threats as 8.1 percent—and mobilization as 6.8 percent—more credible than the baseline condition. Although 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 Hypothesis 2, we find no evidence that one type of signal is interpreted as being significantly weaker than the other, all else equal.

The bottom panel of Figure 2 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 percent and 20.9 percent more credible than the baseline condition, more than double the size of the effect among leaders.6

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 hold that democracies are better able to credibly signal their resolve with public threats than their nondemocratic counterparts, but in accordance with some other recent work (Tomz 2009; Renshon, 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 percent, 2.3 percent less than the same signal sent by a dictatorship, although this difference in difference is not statistically significant (bootstrapped p < .171).7

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 3. The list of experiences studied (see Table 3) 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 in different measures of political experience, whether direct (e.g., experience on the foreign affairs committee or fighting in
Figure 3. Moderating effects of leader characteristics on perceived efficacy of signals. Note: Panel (a) presents the bootstrapped distribution of coefficient estimates from model 7 from Table 2 in Online Appendix §3.3; panel (b) presents the bootstrapped distribution of coefficient estimates from model 7 from Table 3 in Online 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 such as 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.
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. 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 with all the covariates. Our main findings are illustrated in Figure 3, 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).

**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 3. 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 2018). As such, hawks and doves have different reference points: doves view foreign policy through 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 4, 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 8.4 points more when the opponent issues a verbal threat compared to when they mobilize military forces \( (p < .03) \). In fact, supplementary analyses suggest that for the most dovish 22 percent 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.\(^{11}\) 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 3.8 points more in response to mobilization than a public threat \( (p < .12) \). In fact, threats aren’t seen as credible at all by the most hawkish 13 percent of our leaders. As a result, the difference-in-difference-in-difference (or, the difference in the within-subject effects of public threats compared to mobilization between doves and hawks) is substantively large (12.25 points) and statistically significant \( (p < .01) \).

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

<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>8.41 ( (p &lt; .03) )</td>
</tr>
<tr>
<td>... public threat compared to mobilization</td>
<td>Hawk</td>
<td>-3.84 ( (p &lt; .12) )</td>
</tr>
<tr>
<td>Difference-in-difference-in-difference</td>
<td></td>
<td>-12.25 ( (p &lt; .01) )</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).
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 Hypothesis 3b. High-trust leaders view mobilization of troops and public threats as more informative signals of resolve than low-trust leaders do. However, when we compare the relative efficacy of these two types of signals, high-trust leaders tend to view mobilization of troops as a more significant indicator of future behavior than public threats, similar to hawks.

**Leader experiences.** Thus far we have shown 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 3 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 percent 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 Online 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 posttreatment 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 and 5 of Online 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.

**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 overlearning 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, perhaps 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 and stakes) 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 treatment (though see Tomz 2009).

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 (Saunders 2018; Horowitz et al. 2018). However, although our findings about the credibility of public threats are 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 mean 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 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 posttreatment 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 complex 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)? 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 prepolitical 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.
Authors’ Note
A previous version of this article received the Best Paper Award from APSA’s Foreign Policy Analysis section. This is one of several joint articles by the authors; the ordering of names reflects a principle of rotation.

Acknowledgments
The authors would like to 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, Belgin San-Akca, Elizabeth Saunders, Rob Schub, Rachel Stein, Todd Sechser, Jack Snyder, Caitlin Talmadge, Mike Tomz, Jane Vaynman, and audiences at GWU, UC Berkeley, UC Merced, Cornell University, Temple University, Koç University, Emory University, APSA 2016, and Peace Science 2016 for helpful feedback. The authors are also grateful to Lotem Bassan, Roshni Chakraborty, and Michael Sagiv for invaluable research assistance and to Mike Tomz, Jessica Weeks, Roni Milo, Dr. Shirley Avrahami, and Yardena Miller at the Knesset—without whom the elite survey would not have been possible.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material
Supplemental material for this article is available online.

Notes
1. See also Weiss (2013) on how autocrats can use nationalist protests to generate costly signals.
2. In this sense, much of the skepticism about costly signaling in international relations (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.
3. 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.
4. Men 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 Israel Defense Forces.
5. Our focus on the Israeli Jewish population alone in this survey is entirely due 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.
6. The 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 percent and 22.3 percent, respectively). For this section of the analysis, we
average across regime types (democracy and dictatorship) for the public sample.

7. When we screen out participants who failed the manipulation checks, the difference in
difference increases in magnitude to $-3.1$ percent, although the effect is still not signif-
ican (bootstrapped $p < .152$).

8. Because of the ubiquitousness of military service in Israel, 95 percent 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.

9. 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 mak-
ing causal claims about their effects.

10. The complete regression tables are presented in Online Appendix §3.3.

11. 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 1,500 times, and identifying the values of militant
assertiveness for which the 95 percent confidence intervals cross zero.

12. 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 them-
seves are different but rather that they face different incentive structures.

References


Foreign Policy Analysis 6 (3): 257-73.

of Legal Obligation in International Politics.” International Organization 71 (S1):
S136-63.


Yarhi-Milo, Keren. 2014. Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations. Princeton, NJ: Princeton University Press.