Backing Out or Backing In? Commitment and Consistency in Audience Costs Theory

Jack S. Levy   Rutgers University
Michael K. McKoy   Wheaton College
Paul Poast   University of Chicago
Geoffrey P.R. Wallace   Rutgers University

Audience costs theory posits that domestic publics punish leaders for making an external threat and then backing down. One key mechanism driving this punishment involves the value the public places on consistency between their leaders' statements and actions. If true, this mechanism should operate not only when leaders fail to implement threats, but also when they fail to honor promises to stay out of a conflict. We use a survey experiment to examine domestic responses to the president's decision to "back down" from public threats and "back into" foreign conflicts. We find the president loses support in both cases, but suffers more for "backing out" than "backing in." These differential consequences are partially explained by asymmetries in the public's treatment of new information. Our findings strongly suggest that concerns over consistency undergird audience costs theory and that punishment for inconsistency will be incurred, regardless of the leader's initial policy course.

Replication Materials: All data and Replication files are available at http://dx.doi.org/10.7910/DVN/5HEB4I. Any errors remain our own.

In his first inagural address, in the midst of the 1861 secession crisis, Abraham Lincoln declared "there will be no invasion [of the Southern states]—no using of force against or among the people anywhere" (Basler 1953, 254). Following the bombardment of Fort Sumter on April 12, 1861, Lincoln, hoping to reunite the country with minimal bloodshed, imposed a blockade and reaffirmed a policy of noninvasion. As he wrote in a late April letter to Maryland Representative Reverdy Johnson, "I *do* say the sole purpose of bringing troops *here* is to defend the capital. I *do* say I have no purpose to *invade* Virginia, with them or any other troops, as I understand the word *invasion* (emphasis in the original)" (Basler 1953, 343). By the end of June, over two months after the attack on Fort Sumter, Lincoln shifted course. On June 25 and 29, Lincoln summoned two war council meetings to authorize an offensive strike against Southern forces at Manassas Junction, Virginia (Poast 2015). Union forces subsequently lost the Battle of Manassas (Bull Run)

on July 21, 1861, which compelled Lincoln to sign into law the raising of half a million soldiers.

Lincoln broke his promise not to invade the South. Lincoln was not the first leader—and surely not the last—to make a public commitment during a crisis and then subsequently break that promise. Yet, what makes Lincoln's promise noteworthy—particularly given the extensive literature on "audience costs" that focuses on the consequences of a leader's public threat to *use* military force—is that Lincoln made a public promise to *abstain* from using military force. Lincoln's behavior does not appear to fit the assumptions of audience costs theory, which argues that the public values consistency between the statements and actions of leaders and fears the reputational costs of making threats but not following through (Fearon 1994a; Guisinger and Smith 2002; Schultz 2001, 2012; Slantchev 2006; Smith 1998; Snyder and Borghard 2011; Tomz 2007; Weeks 2008). Hence, the theory predicts that a leader's anticipation of domestic political costs for

inaction increases the likelihood that she or he will follow through on the threat, enhancing the threat's credibility. However, if the public values consistency between words and deeds, should we not also expect the public to similarly punish leaders for failing to stand by a public promise to refrain from action?

We argue that if the domestic public punishes a leader for inconsistency, then the public should punish a leader for failing to uphold a promise to refrain from action, not just for backing down from a threat to use force. This argument, and our approach to evaluating this argument, enable us to test the centrality of consistency as a crucial mechanism for audience costs. Building upon Tomz's (2007) groundbreaking study, we use a survey experiment to examine domestic responses to the president's decision to back down after issuing a public threat or to use force after promising to stay out of a conflict. We find that the leader incurs a loss of support for refusing to follow through on a previous threat ("backing out") or when they militarily engage abroad despite earlier public promises to abstain from using force ("backing in"). Furthermore, the leader suffers a greater loss of support for "backing out" than "backing in."

These differential consequences are explained in part by an asymmetry in the public's treatment of new information. Following Levendusky and Horowitz (2012), we included an additional treatment regarding whether the president stipulated that new information motivated the change in policy. Like Levendusky and Horowitz, we find that new information mitigated the punishment associated with inconsistency between words and deeds.[1] However, the mitigating effect is larger for "backing out" than for "backing in." We further show that concerns over the credibility of future U.S. promises figure prominently in understanding the public's punishment of both forms of inconsistency, while evaluations of leadership competence and broader reputational matters have more limited and uneven mediating effects. Overall, our findings strongly suggest that concerns over consistency undergird audience costs theory, and that punishment for inconsistency will be incurred, though to varying degrees, regardless of the initial policy course chosen by the leader.

Evaluating Consistency in Audience Costs Theory

Building on earlier discussions of the domestic consequences of external threats (Jervis 1970; Schelling 1960),

Fearon (1994a, 1997) developed and formalized the concept of "audience costs." The potential for domestic audience costs is generated when a leader makes a threat that is observed by a domestic public possessing the power to punish the leader for failing to abide by the commitment.[2] These costs can include lower popularity, greater domestic opposition to policy initiatives, removal from office, and in worst-case scenarios, imprisonment or execution (Goemans 2008). Fearon (1994a, 581) specifically argues that domestic concerns about the "*international* loss of credibility, face, or honor (emphasis in original)" would increase the costs to a leader for backing down after publicly issuing a threat. These additional costs for backing down—which are distinct from the public's evaluation of the issue in dispute—make the threat more credible. Guisinger and Smith (2002) further argue that domestic audiences, recognizing the necessity of an honest reputation in international negotiations, punish leaders who damage the country's reputation by pursuing policies in contradiction with public pronouncements. Though first conceptualized in regards to military threats, audience costs as a commitment mechanism have since been applied to international negotiations more broadly and across a wide range of issue areas (Broz 2002; Jensen 2003; Martin 2000; Simmons 2010).[3]

Nevertheless, there remains debate about the existence of audience costs and their motivating microfoundations (Chaudoin 2014; Levy 2012; Sartori 2005; Slantchev 2006; Smith 1998; Snyder and Borghard 2011; Trachtenberg 2012). Unfortunately, much of this earlier empirical work was stymied by selection effects, whereby leaders strategically choose to generate or avoid the presence of audience costs.[4] Tomz (2007), seeking to address these issues, employed a survey experiment and found that participants responded very unfavorably when the president failed to follow through on a threat. In particular, 72% of respondents said their disapproval was based on the president's inconsistent behavior, "by saying one thing and doing another."[5]

---

[2] Leaders also face reputational costs from external audiences, but we follow the literature and focus on domestic audiences.

[3] Uzonyi, Souva, and Golder (2012, 768) broaden audience costs to be any instance of the public punishing a leader for poor foreign policy choices. We follow the more standard definition of restricting audience costs to the domestic costs for failing to follow through on an explicit threat or commitment.

[4] If leaders anticipate that a particular threat will generate large audience costs should they subsequently have to back down, they will be less likely to make the threat, and large audience costs will rarely be observed (Schultz 2001).

[5] That percentage drops to 61% if one excludes those who viewed the president as incompetent for apparently misreading the situation.

---

[1] We actually find that participants *rewarded* the president for changing course after citing new information in the "backing out" scenario.

Each of the hypothesized microfoundations of audience costs theory–including reputations for resolve or honesty, and perceptions of the leader's competence– presumes consistency between leaders' commitments and actions.[6] Most attempts to evaluate the consistency mechanism have concentrated almost entirely on situations in which leaders make coercive threats and promise to act if an aggressor fails to comply with the terms of the threat. This exclusive focus on such "positive" commitments is understandable, given the standard definition of audience costs as "the domestic price that a leader would pay for making foreign threats and then backing down" (Tomz 2007, 821; see also Fearon 1994a, 581; Schultz 2001, 33). However, if domestic punishment for backing down from a threat is driven, even in part, by a distaste for inconsistency, then the punishment should be incurred not only when leaders fail to follow through on a threat, but also when they fail to honor public commitments to stay out of a conflict. These latter scenarios can be thought of as decisions to renege on "negative commitments," where a leader makes a promise *not* to engage in a particular course of action. The consistency mechanism of audience costs theory generates observable implications for violating negative commitments (e.g., promises to stay out) as well as for violating positive commitments (e.g., promises to intervene if the adversary does not change its behavior). Systematically testing the consistency mechanism requires looking both at situations in which leaders make a threat but then back down, and situations in which leaders publicly promise not to act militarily but then decide to act. Leaders should thus be punished for inconsistency regardless of whether they "back out of" or "back into" conflicts.

Figures 1 and 2 help illustrate our argument. Figure 1 shows a standard decision tree in a crisis bargaining game commonly used in previous research, particularly in the audience cost literature.[7] For ease of presentation, we use the United States as the sender and a generic foreign country as the target. In this game, the leader can choose whether to issue a threat (the "Threaten" decision) in response to an unwanted action by another actor, such as the commonly used scenario of a foreign country invading its neighbor. If the leader issues a threat, and if the target does not concede to the threat and continues its aggression, the leader then decides whether to actually use military force. If the leader does not follow through

on the threat and refuses to deploy military force, then the game ends in the "Back out" outcome. If the leader instead follows through and deploys force, then the game ends in the "Go in" outcome.

Alternatively, if the leader chooses not to issue any threat in the first place (the "Don't threaten" decision) and instead promises that their country will stay out of the conflict, then the game automatically ends with the "Stay out" outcome. We should note that the commonly used label "Don't threaten" to refer to the initial decision on the left-hand branch involves some ambiguity, though we employ it for the purposes of Figure 1 to reflect the standard understanding of the crisis bargaining model. The existing experimental literature on audience costs generally views this action as referring to a public statement that the leader's country will stay out of the conflict (Tomz 2007, 824). However, an alternative possibility is that the leader simply remains silent and makes no public statement, threatening or otherwise. The latter might generate a different response if the leader subsequently resorts to military force. Given our theoretical focus on consistency between words and actions, in our own later model we use the label "Promise to stay out" to reflect an explicit negative commitment.
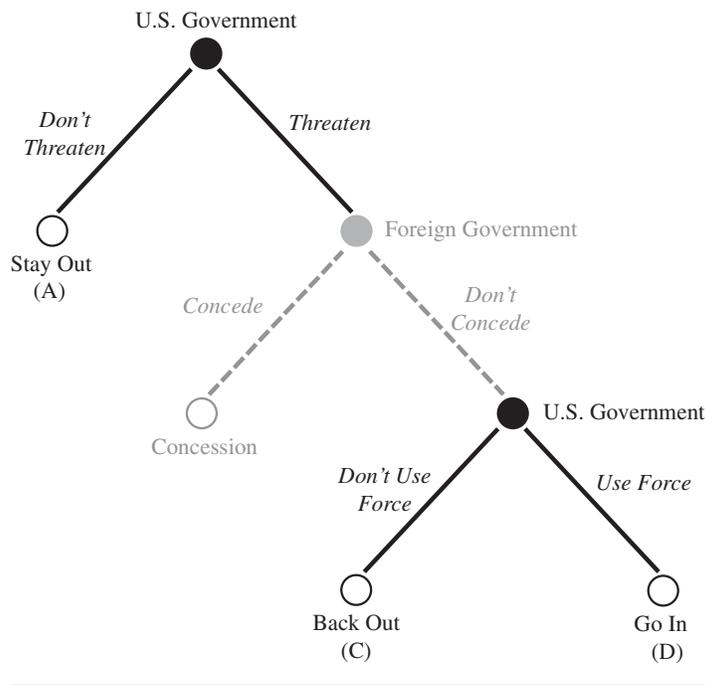
Existing studies have focused on and added a variety of features to the right-hand "Threaten" side of the game tree in Figure 1—such as whether the foreign country concedes—while designating "Stay out" as the baseline condition on the left-hand side (Brutger 2014; Kertzer and Brutger, forthcoming; Levendusky and Horowitz 2012; Tomz 2007; Trager and Vavreck 2011). However, treating "Stay out" as a terminal node is problematic both theoretically and empirically. Abraham Lincoln remained committed to not invading the Confederacy, even after the attack on Fort Sumter, but later relented. Woodrow Wilson campaigned on keeping the United States out of World War I, despite several earlier German naval provocations. Nevertheless, he ultimately chose to declare war on Germany in March 1917 (Devlin 1974).

Such examples simply emphasize that, depending on the theoretical question under consideration, it is useful to treat the node resulting from the "Promise to stay out" decision as a decision node rather than a terminal node. This is depicted in our revised crisis bargaining game in Figure 2. After publicly promising that the country will stay out of the conflict, the leader has the option of either honoring this negative commitment, or reversing course and using force. If the leader sticks to the negative commitment, the country continues to stay out of the conflict. If the leader instead decides to forcefully respond to the foreign country's initial provocation, the state intervenes in the conflict, which we label as the "Back in" outcome.

---

[6]Tomz (2007, 835–836) found a mix of reasons for disapproval, though most respondents were concerned about the state's international reputation for resolve, while a minority mentioned a normative preference for honesty.

[7]Scholars vary in the specific labels they attach to the various decisions and outcomes.

FIGURE 1 Standard Crisis Bargaining Model

U.S. Government

Don't Threaten   *Threaten*

Stay Out
(A)

Foreign Government

Concede   *Don't Concede*

Concession

U.S. Government

Don't Use Force   *Use Force*

Back Out
(C)

Go In
(D)

Like the "Back out" outcome, it involves a reversal of the leader's initial stance, and constitutes an inconsistency between commitments and actions.

Converting the "Stay out" terminal node into a decision node brings into focus how either policy commitment–"Threaten" or "Promise to stay out"–is now subject to defection. If the president publically promises to stay out of a conflict rather than issue a threat, then this is equivalent to committing to non-involvement. The leader's statements set the public's expectations regarding the leader's future actions. If the public values consistency between the leader's words and deeds, the public will punish the leader for committing to stay out of the conflict but then deciding to intervene. Hence, extending the game tree beyond the initial left-hand decision to stay out of the conflict provides insight into the importance of consistency and the domestic consequences that can result for leaders. If domestic audiences value consistency or view a broken commitment as a sign of incompetence (Smith 1998), then the public should punish leaders both when they make a threat but fail to follow through *and* when they state that they will stay out of a conflict but then intervene. If consistency is important, each of these scenarios should generate "inconsistency costs" from domestic publics. As Jensen (2003, 595) writes in the context of honoring agreements to protect foreign investments, "democratic leaders are held accountable for their actions, including reneging on a *promise* or threat (emphasis added)."

Distinguishing between these two kinds of inconsistency costs raises the question of whether there are theoretical reasons for expecting these costs to be equal in magnitude, or whether the public punishment for "backing out" is greater (or less) than the punishment for "backing in." Beyond their exact size and direction, the public may also weigh concerns over consequences to their country's reputation and credibility differently when evaluating the fulfillment of each type of commitment. The international relations literature provides little guidance on this matter. Scholars have given a lot of attention to the credibility of threats, but almost none to the credibility of promises.[8] Little has changed in this respect since Jervis (1991, 25) wrote "almost nothing has been said about the value of a reputation for living up to promises."[9] Yet, even the limited discussion of promises has little direct relevance for our question. Threats and promises are commonly defined in terms of contingent deprivations and improvements on a target's value position (Davis 2000, 12). This has no clear implications for the evaluations of violations of threats and promises by domestic publics. There is a more substantial literature

[8]Scholars generally recognize that threats and promises are not fully separable, that each is a "selective and conditional self-commitment" (Schelling 1960, 134).

[9]Similarly, Davis (2000, 2) wrote, "there exists little theoretical or empirical research on the use of promises in international relations."

**FIGURE 2   Revised Crisis Bargaining Model with "Promise to Stay Out" Leading
to Decision Node**
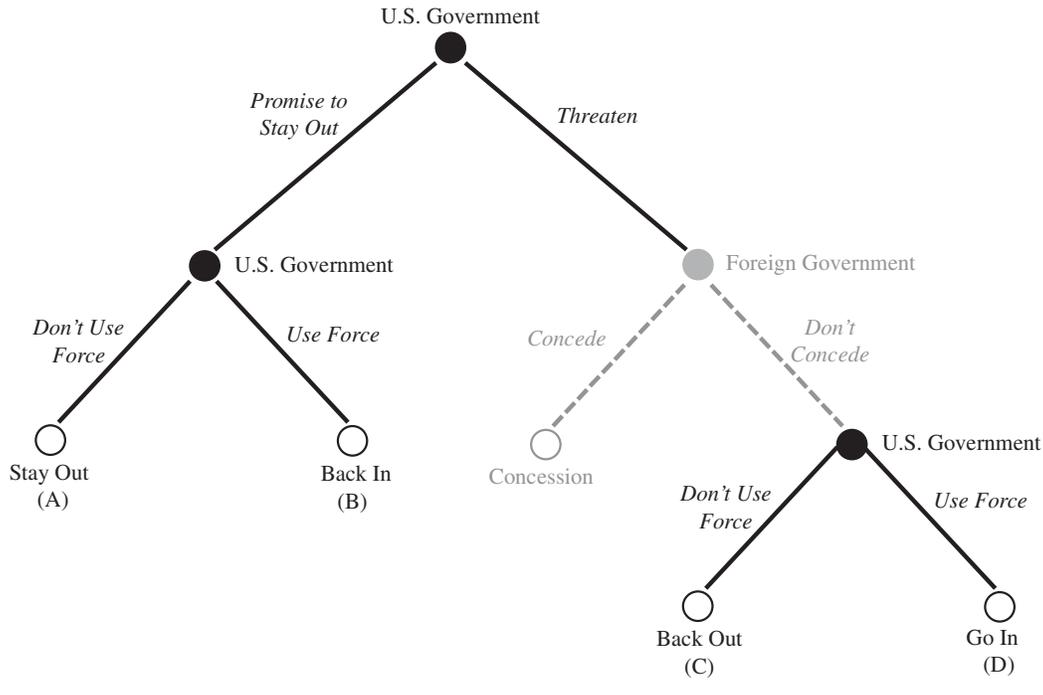
comparing threats and promises in social psychology, but it too focuses on the future dyadic relationship between initiator and target, and says nothing about accountability to a third party.[10]

In addition, in our experiments the "promise to stay out" is not a promise in the same sense as in the social psychology literature, where a promise offers a positive reward to a target conditional upon the target's behavior (Davis 2000). The "promise" in our experiment does not offer a positive reward, but only the absence of a conditional punishment. In addition, the promise is not explicitly contingent on any particular behavior of the target,[11] and it is directed toward domestic publics at least as much as toward an external target. Given this lack of theoretical guidance, it is best to treat the direction and magnitude of any asymmetries in the consequences of reneging on positive and negative commitments as an empirical question.

Data and Methods

Our design follows the general setup of the survey experiment originally outlined in Tomz (2007), but with several important modifications described below. To ensure any differences in results are not simply due to changes in the instrument, we used the same scenario and phrasing to the greatest extent possible. The survey begins with a brief introductory script about leaders and how they handle different foreign policy situations. Next, the survey informs respondents that they will be asked to evaluate one approach that U.S. leaders have taken in such situations. All respondents are then presented with a crisis where a foreign country used military force to invade a neighboring country.[12]

Participants were randomly assigned to one of six experimental groups based on three separate treatments. The various groups, along with the corresponding number of respondents, are summarized in Table 1. A series of balance tests indicates that groups are comparable across

[10]One leading line of argument in social psychology is that, at the dyadic level, promises are more credible than threats (Heilman 1974; Rubin and Lewicki 1973).

[11]For a comparison of the effects of contingent and non-contingent threats and promises, see Cheney, Harford, and Solomon (1972).

[12]Tomz randomly varied several contextual factors (regime type, motive, power, and interests) to guard against idiosyncratic features of the crisis driving the findings. To maximize the statistical power of our results and reduce concerns over possible interactions among the various treatments, we do not vary these contextual factors. This is similar to the strategy adopted by several subsequent studies (e.g., Levendusky and Horowitz 2012; Trager and Vavrek 2011).

TABLE 1 Experimental Groups

Group	Threat	Use of Force	New Information	Number of Respondents
Stay Out	No	No	–	369
Back In	No	Yes	No	372
Back In/New Info	No	Yes	Yes	370
Back Out	Yes	No	No	372
Back Out/New Info	Yes	No	Yes	378
Go In	Yes	Yes	–	365

Note: For the Threat treatment, the "No" entry refers to the initial "Promise to stay out" condition illustrated earlier in Figure 2.

a variety of sociodemographic and political traits that could affect foreign policy preferences or evaluations of the president.[13] We now describe the assignment of the three treatments.

The first treatment concerned whether the U.S. president made an initial threat to intervene against the attacking country. This treatment divides respondents into two groups. The control group is told that the president "said the United States would stay out of the conflict." Respondents in the second group are instead told that the U.S. president made an initial threat, announcing "if the attack continued, the United States would send military forces to help to push out the attacking country." The treatment thus corresponds to the verbal threat treatment from the original Tomz (2007) study. While Tomz also varied the level of escalation in the threat to include displays or limited uses of force, we chose to use a simpler binary treatment to maximize statistical power and focus more clearly on our main dynamics of interest concerning the consistency of commitments. This is similar to other studies, such as Levendusky and Horowitz (2012) and Trager and Vavreck (2011), which do not vary the level of escalation in the president's initial threat.[14]

The second treatment examines whether the U.S. president then chooses to ultimately employ military force to help to push out the attacking country. Respondents in both groups—those told that the president issued a threat and those told that the president said the United States will stay out of the conflict—are then informed that the attacking country continues to invade. Respondents originally told that the president "said the United States would stay out of the conflict" are further divided: Some are told that the president did not send the military, while others are told the president deployed armed forces.

Respondents originally told that the president announced that "if the attack continued, the United States would send military forces to help to push out the attacking country" are divided in a similar manner: Some are told that the president did not send the military, while others are told the president deployed armed forces. Other than the decision over using military force, we used the same general language for both groups and avoided terms like "backed down" or "gave up," which may have negative connotations that could bias responses. Similarly, in the conditions where the United States eventually deployed military forces, we avoided terms like "followed through," "stood firm," or "pushed ahead," which could correspondingly positively bias respondents in favor of the president's course of action.

Allowing the president to follow through or back down regardless of the initial policy decision further differentiates our design from Tomz (2007) and several subsequent studies (e.g., Davies and Johns 2013; Levendusky and Horowitz 2012). These other survey experiments do not vary this stage of the scenario; in all instances, the president backs down and does not use force. Other works (Kertzer and Brutger, forthcoming; Trager and Vavreck 2011) include situations where the United States uses military force, but they only do so for the right-hand side of the crisis bargaining game where the president made an initial threat. By including the use of military force even when the president initially said the United States would stay out of the conflict, our design allows for a more complete examination of how consistency between commitments and subsequent behavior influences public approval or punishment of the president's crisis bargaining behavior. To focus the analysis on evaluations of the crisis bargaining actions or non-actions of the president, we leave the ultimate consequences of those decisions unspecified. That is, the final outcomes are the non-use of force and the use of force, not the consequences of each of those actions (degree of success or failure, number of casualties, etc.).

[13]Full results available from the authors upon request.

[14]See Davies and Johns (2013) for one example that does vary the level escalation.

The third treatment builds on the work of Levendusky and Horowitz (2012) to assess whether justifications made by the president based on new information mitigate the negative effects of changing course. Unlike the first two treatments, the new information treatment is only applied to the situations involving inconsistent behavior between the president's initial policy and subsequent decision over the use of force—in this case, either (1) the "Back out" condition where the president pulled back from using force after making an initial threat to do so, or (2) the "Back in" condition where the president eventually turned around and deployed military force after initially saying the United States would stay out of the conflict. Across both instances of inconsistent behavior, respondents assigned to the control were given no additional prompt. Respondents assigned the treatment were instead told that the president received new information suggesting that, for those in the "Back in" condition, involvement in the crisis was in U.S. interests and military experts agreed the United States should become involved. For those in the "Back out" condition, the new information instead suggested that involvement in the crisis was not in U.S. interests and military experts agreed the United States should not become involved.

After being presented with their particular scenario, respondents were then asked, as they were in Tomz (2007) and subsequent studies, to rate their approval of the way the president handled the situation.[15] Responses were combined into a seven-point Likert scale ranging between approving very strongly to disapproving very strongly.[16] The full instrument is included in Supplementary Appendix A.

---

[15]We also included an earlier item immediately after the first threat treatment to measure respondents' initial approval of the president's decision to either make a threat or announce the United States would stay out of the conflict. However, since the core logic of audience costs theory centers on whether states follow through on commitments, we focus on the subsequent approval for how the president handled the overall situation. While asking this initial approval item may bias in favor of finding costs for inconsistency by anchoring respondents to their views on the president's initial reaction, there is little reason to expect the size of the bias to be greater for positive versus negative commitments. A subsequent round of the survey on a smaller sample excluded the prior item on initial approval (the instrument is described in Supplementary Appendix D). Overall results remained substantially the same (see Supplementary Appendix E).

[16]Participants were first asked whether they approve, disapprove, or neither approve nor disapprove of the president's actions. Respondents who answered either "approve" or "disapprove" were then asked a follow-up question to assess whether their opinion was very strong or only somewhat strong. Those answering "neither" were instead asked if they leaned toward approving, disapproving, or neither way.

We fielded our survey from June through August 2014 and recruited 2226 respondents online using Amazon's Mechanical Turk (mTurk) service. mTurk provides access to a recruitment pool of respondents by promising compensation for completing a particular task—in this case, taking an online survey. The advantage of using mTurk in a study of public opinion is that it is a very efficient way of administering surveys without sacrificing much in terms of representativeness. Berinsky et al. (2012) show that subjects recruited on mTurk are more representative of the U.S. population than other common convenience samples, though somewhat less representative than subjects recruited via national probability samples. They replicate several existing studies using subjects recruited through mTurk and find results that are comparable to those produced with other subject pools.

Our respondent pool was relatively close to nationally representative surveys, though unsurprisingly for an Internet-based sample, respondents tended to be younger, predominantly male, and more educated compared to national benchmarks.[17] This is generally consistent with the mTurk-based survey pool acquired by Chaudoin (2014) and others.[18] To verify that these characteristics do not moderate the effects of our treatments, we re-ran our main analysis on various sociodemographic subsamples, such as by gender or level of education. Doing so produces results similar to those reported in the main text, which further suggests our findings are unlikely to be a function of any particularities in the composition of our sample.[19]

Analysis

Because of randomization, complex statistical models involving a battery of covariates are unnecessary for obtaining internally valid inferences regarding the effects of crisis bargaining behavior by the president on domestic approval. Following Tomz and several other experimental studies, we report simple cross-tabulations for the

---

[17]Benchmarks for the U.S. adult population were based on the June through August 2014 updates of the Current Population Survey (CPS). Several key sociodemographic variables for the mTurk sample compare to CPS benchmarks as follows: 58% versus 48% of respondents were male; 86% versus 44% were below the age of 44 years; 47% versus 30% had completed at least a college degree. Full results are available in Supplementary Appendix B.

[18]See, for instance, Bernauer, Gampfer, and Kachi (2014).

[19]All coefficients remain of comparable size and in the same direction, though in some instances the statistical significance is attenuated, in part due to the smaller size of the subsamples. See Supplementary Appendix C.

TABLE 2(a) Domestic Political Consequences of Being Inconsistent (Outcome = No Force)

% Who Approve Continuing to Stay Out("Stay Out")	% Who Approve Not Following Through on Issued Threat("Back Out")	*Extent that Issuing Empty Promise Hurt (Difference in approval)*
51 (46, 57)	29 (24, 33)	*–22 (–16, –30)*
(n = 369)	(n = 372)	

TABLE 2(b) Domestic Political Consequences of Being Inconsistent (Outcome = Use Force)

% Who Approve Following Through on Issued Threat("Go In")	% Who Approve Not Continuing to Stay Out("Back In")	*Extent that Issuing Empty Promise Hurt (Difference in approval)*
53 (47, 58)	41 (36, 46)	*–12 (–4, –19)*
(n = 365)	(n = 372)	

Note: 95% confidence intervals in parentheses. Reported first differences may differ slightly due to rounding.

relevant values of interest.[20] Table 2(a) displays three values: (1) the percentage of respondents, told that the president decided to stay out of the conflict, who strongly or somewhat approve of the president's decision; (2) the percentage of respondents, told the president made a verbal threat and subsequently backed down, who strongly or somewhat approve of the president's decision; and (3) the difference between the two percentages. We compare the "Stay Out" condition with the "Back Out" condition, as the ultimate outcome is the same in both scenarios (no force is used), allowing us to isolate the "punishment" associated with inconsistency in statements. The 22-percentage point fall in presidential approval suggests that leaders do experience a noticeable drop in approval for being inconsistent.[21] Indeed, the 22-percentage point drop is, in fact, larger than the 16% penalty observed in the original Tomz study (though our confidence interval does contain the original 16% change found in Tomz).[22] Hence, this first result is in line with prior experimental work and is consistent with the overall audience costs logic: Domestic publics seem to punish leaders for inconsistency between words and deeds when not following through on their initial commitments.

[20]Subsequent analyses (see Supplementary Appendix C) show the results are robust to the inclusion of a number of standard sociodemographic variables.

[21]All analysis conducted using Stata 13. Some reported effects diverge slightly from the differences between the percentages displayed for the relevant experimental conditions due to rounding. As a robustness check, we also computed the confidence intervals for the first differences using a bootstrap procedure of 1,000 with-replacement resamples of our data. The confidence intervals produced by this procedure (-16, -30) are nearly identical to those reported in the article (code to compute the confidence intervals is provided in the replication packet).

[22]See Table 1 in Tomz (2007, 827).

Table 2(b) is similar to 2(a), except it focuses on the two scenarios that led to the use of force. More precisely, Table 2(b) displays three values: (1) the percentage of respondents, told that the president made and followed through on a verbal threat to use force, who strongly or somewhat approve of the president's decision; (2) the percentage of respondents, told the president initially decided to stay out of the conflict but subsequently chose to use force, who strongly or somewhat approve of the president's decision; and (3) the difference in the two percentages. As with Table 2(a), since the ultimate outcome is the same in both the "Back In" and "Go In" conditions (force is used), the difference in the approval ratings again captures the "punishment" associated with inconsistency in statements. The assumption behind conventional versions of the audience cost argument is that domestic publics value consistency in behavior because empty threats undermine their country's reputation on the international stage or signal their leader's incompetence (Fearon 1994a, 580; Smith 1998; Tomz 2007, 823). If consistency between words and deeds is indeed prioritized above all else, then citizens should punish the leader for "backing into" a conflict as well as for "backing out" of a conflict. The drop of 12 percentage points in presidential approval again suggests that domestic publics seem to punish leaders for not matching their words with deeds.

Note that the decline in approval reported in Table 2(b) is smaller in magnitude than that reported in Table 2(a) (12 percentage points compared to 22 percentage points).[23] Earlier, we had raised the possibility that punishments for "backing out" and for "backing in" might be different, but we concluded that there were no

[23]This difference is confirmed to be statistically significant at the 0.95 confidence level in a two-sided t-test.

TABLE 3(a) Domestic Political Consequences of Being Inconsistent When Citing New Information (Outcome = No Force)

% Who Approve Continuing to Stay Out("Stay Out")	% Who Approve Not Following Through on Issued Threat With New Information("Back Out")	*Extent that Issuing Empty Promise Hurt (Difference in approval)*
51 (46, 57)	65 (61, 70)	*+14 (+7, +21)*
(n = 369)	(n = 378)	

TABLE 3(b) Domestic Political Consequences of Being Inconsistent When Citing New Information (Outcome = Use Force)

% Who Approve Following Throughon Issued Threat("Go In")	% Who Approve Not Continuing to Stay Out With New Information("Back In")	*Extent that Issuing Empty Promise Hurt (Difference in approval)*
53 (47, 58)	51 (46, 56)	*–2 (–9, +6)*
(n = 365)	(n = 370)	

Note: 95% confidence intervals in parentheses.

obvious theoretical reasons for predicting which direction this asymmetry might take, much less its magnitude. Nevertheless, our experimental design allows us to gain some leverage on explaining this asymmetry in punishment. More precisely, to help understand the greater punishment for backing out, we manipulate information as a moderating variable. Consider Tables 3(a) and 3(b). Table 3(a) presents the same information as Table 2(a), with one important exception: Following Levendusky and Horowitz (2012), respondents are told that before choosing not to following through on the threat, the leader received new information indicating the conflict was not in U.S. interests.[24] Rather than being punished, the 14-percentage point increase in approval reported in the third column of Table 3(a) shows that leaders are actually *rewarded* for not following through on a threat if they claim to have received new information. This is consistent with

Levendusky and Horowitz (2012), who also find that appearing to make a prudent decision results in higher approval than a decision to never enter the conflict in the first place.[25]

Compare this result to that in Table 3(b). Table 3(b) presents the same information as Table 2(b), except that respondents are told that the leader received new information when choosing to renege on the prior promise to stay out of the conflict. In this scenario, there is no statistically discernible change in the leader's approval rating when comparing the "Go in" first column (53% approval) and "Back in with new information" second column (51% approval) of Table 3(b). Of course, when the second column of Table 3(b) is compared to the second column (41% approval) of Table 2(b), the change in approval due to new information is still notable (though a more modest rise of 10 percentage points).[26]

These findings present a puzzle: Why does new information result in the public rewarding a president who backs out after making a threat, but does less to change approval for a president who backs in after saying the country will stay out? One possible explanation for the observed response is that respondents make certain assumptions about information, quite apart from the

[24]Strictly speaking, the "New Information" treatment contains two separate components. One element is whether or not the president received new intelligence suggesting involvement was (for "Back in" groups), or was not (for "Back out" groups), in the interest of the United States. The second element is whether or not military experts agreed that the United States should (for "Back in") or should not (for "Back out") become involved in the crisis. We chose this language to offer the highest degree of comparability to the earlier study of Levendusky and Horowitz (2012), but this treatment cannot disentangle whether the relevant effects are driven by new intelligence or the support of military experts. In a follow-up survey, we thus separated the original "New Information" treatment into three separate conditions: "New Intelligence Only," "Expert Consensus Only," and "New Intelligence and Expert Consensus," where the latter exactly mirrors the language from the "New Information" treatment used by Levendusky and Horowitz (2012) and our original survey. The results generally show that all three conditions mitigate the costs of inconsistency, though with a slightly more marked effect for new intelligence over the consensus of military experts (see Supplementary Appendix E).

[25]In Levendusky and Horowitz (2012), approval for never entering is 33%, while approval rises to 39% for a leader who issues a threat, but then backs down after claiming to have received new information. However, in their study the first difference does not achieve standard levels of statistical significance.

[26]The difference in these percentages is statistically significant at the 0.99 confidence level in a two-sided t-test.

treatment conditions in the experiment. In a partial test of this possibility, we asked respondents in the baseline "Back out" and "Back in" conditions (i.e., those groups that did not receive the new information prompt) about their beliefs over whether the president actually received new information that led the president to act inconsistently with the prior declaration (issuing a threat, or saying the United States would stay out, respectively). As expected, respondents were significantly more likely to believe the president obtained and acted on new information in the "Back in" versus "Back out" conditions.[27]

This, of course, pushes the puzzle back a step. How do we explain the asymmetry in respondents' informational assumptions in the two scenarios? One possibility is that the attenuation of domestic punishment associated with the "Back in" scenario in Table 2(a) is due to respondents presuming that the president received new information when the president chose to intervene after first intending to stay out. These asymmetric responses to the president's decision to "back out" and "back in" based on new information are both interesting and surprising, and should be the focus of further research.

Another possible explanation for the greater public punishment for reneging on threats to "Go in" than for reneging on promises to "Stay out" is that the threat involves a more explicit specification of the action(s) that would trigger the implementation of the threatened response. This makes it easier for the domestic audience to identify and then punish inconsistent behavior. In contrast, the promise to stay out in this experiment is not coupled with specific conditions under which the promise will be honored. It is possible that subjects might impose their own imagined conditions. One possibility is that subjects assume that the promise to stay out is unconditional. Another possibility, which we think is more likely, is that subjects interpret a promise to stay out to include the qualifier "at this time." If so, a subsequent decision to reverse course and intervene would not necessarily be interpreted as blatantly inconsistent with the initial commitment. This line of argument could be tested by manipulating the wording of the "promise to stay out" condition to indicate varying levels of consistency in the president's promise, but we save that for another study.

[27]The difference in the proportion of respondents between the "Back in" and "Back out" conditions who answered either "very likely" or "somewhat likely" that the president received new information was 8% with 95% confidence interval of (2, 14).

Unpacking the Consequences of Inconsistency

In his original article, Fearon (1994a, 581 fn.12) posited a number of underlying motives accounting for the audience costs that result from a leader backing down in the midst of a foreign policy crisis. We thus asked a series of follow-up questions to further probe the public's rationale for disapproving of the president's inconsistency across each type of commitment. For instance, Levendusky and Horowitz (2012, 334) show that public perceptions of the leader's competence figure prominently in the decision to punish his or her actions. We therefore asked subjects to rate the competence of the president on a five-point scale ranging from very competent to very incompetent. Similarly, because audience costs theory attaches great importance to the reputational harm purportedly resulting from failing to follow through on promises (Tomz 2007, 835–836), respondents were also asked how much "the president's handling of the situation helped or hurt the reputation of the United States in the world" on a corresponding five-point scale. Of course, the public might have different understandings than do scholars and practitioners of the meaning of "reputation," or may infer that different sorts of reputation may be on the line depending on the situation. The most common reputation of concern in standard models of crisis bargaining focuses on a country's credibility to carry out its threats and promises (Fearon 1994a; Schelling 1966, 124). We thus ask an additional question focusing on respondents' perceptions of how the president's behavior in the crisis affects the likelihood "other countries would believe threats and promises made by the U.S. president in the future" on a four-point scale ranging from "very likely" to "very unlikely."

Figures 3(a) through 3(c) summarize the core findings from these follow-up questions. Each subfigure plots four points for each of the rationales that might explain the punishment for inconsistency. The first two points in each subfigure report the first differences for inconsistency between words and deeds for both the positive commitment (where the president backed down from an initial threat) and the negative commitment (where the president backed in and deployed military force after originally saying the United States would stay out of the conflict). The second set of points in each subfigure reports the first differences for inconsistency between words and deeds when the president cited new information for justifying the inconsistent behavior (be it for a positive or negative commitment). As with the main results in Tables 2 and 3, the baseline category for comparison is kept the same for positive commitments on the one hand

FIGURE 3 Potential Motives Underlying the Consequences of Inconsistency
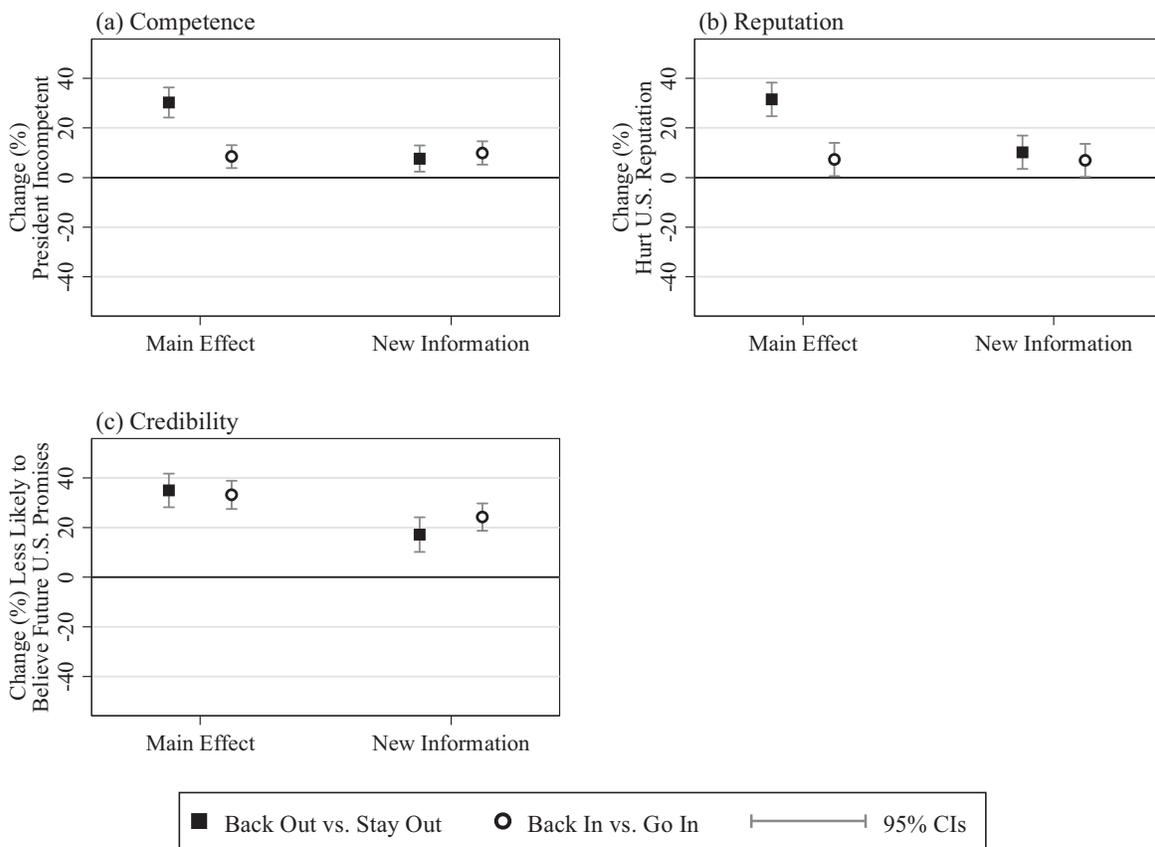
Note: Values display first differences for inconsistency between words and deeds for the relevant type of commitment, separated by whether or not the president cited new information in justifying his or her behavior. First differences report the change in percent answering the following answer options for each outcome variable: President's competence (very incompetent / somewhat incompetent); U.S. reputation (hurt a lot / hurt somewhat); Credibility for likelihood of other countries believing future U.S. promises (very unlikely / somewhat unlikely). All alternative outcomes are scaled such that higher values indicate greater negative consequences for the president.

("Stay out"), and negative commitments on the other ("Go in"). Vertical lines report 95% confidence intervals for each treatment effect. For further ease of interpretation, all of the follow-up outcomes are scaled so that higher values indicate greater negative consequences for the president in terms of his or her competence, the reputation of the United States, or the credibility of future U.S. promises.[28]

The figures show that backing down from an initial threat carries significant costs across several of these related dimensions. Compared to staying out of the

conflict completely, failing to follow through on a threat is associated with more than a 30-percentage point rise in the likelihood that a respondent views the president as incompetent (Figure 3[a]). This is also the case for respondents who feel that failing to follow through on a threat hurts U.S. reputation (Figure 3[b]) and for respondents who believe that future threats and promises made by the president will be less credible (Figure 3[c]). However, as we found in Table 3, justifying the decision to back down by referring to new information diminishes the repercussions to leaders across these three areas, especially for competence and reputation.

By contrast, the dynamics for failing to follow through on negative commitments demonstrate some remarkable differences from failing to follow through on positive commitments. Since positive commitments have

---

[28] As with the main presidential approval outcome measures, the results for these follow-up outcomes remain substantially the same when estimating separate regression models, including the various treatment variables and a number of standard sociodemographic covariates.

been the traditional focus of crisis bargaining models, these results further support the merit of evaluating when leaders decide to turn around on a prior promise to stay out of a conflict. Concerns over the competence of the president are much weaker for negative commitments compared to positive commitments: There is only an increase of 8% that the respondent will view the leader as incompetent (Figure 3[a]). There also appears to be little concern for reputational consequences, broadly understood by the public, due to being inconsistent on a negative commitment (Figure 3[b]).[29] Where leaders do seem to be hurt is regarding concerns about the credibility of future U.S. promises (Figure 3[c]). When backing into a conflict in which the president had previously committed to stay out, there is a more than a 30-percentage point increase in the likelihood that respondents believe future U.S. promises will lack credibility. This effect is commensurate with that found for backing down from a prior threat. Unlike in the case of backing down, however, justifying the decision to back into a conflict by citing new information does not significantly reduce the consequences of inconsistency across the competence or reputation items. This is likely because the negative effects on competence and reputation are already quite small. It does appear that new information has a modest mitigating effect on perceptions of credibility. The general lack of any effect for new information for reneging on negative commitments is consistent with our prior main results for presidential approval (see Table 3).

These follow-up results should be interpreted cautiously and are best viewed as a first cut, given possible conceptual overlap and respondents' potential differential understandings of these various items, which are meant to tap into the various dimensions of domestic punishment.[30] Nevertheless, taken together, the results in Figure 3 suggest that worries over the credibility of the promises

and threats made by the United States down the road appear to figure prominently in explaining the consequences of inconsistent crisis bargaining behavior. The results for credibility hold irrespective of the type of commitment–positive or negative–breached by the leader. On the other hand, evaluations of the president's competence or more general concerns about U.S. reputation are much more apparent for "backing out of," rather than "backing into," a conflict during a crisis.

Conclusion

Audience costs have been a critical, yet controversial, component of theories of international crisis bargaining since Fearon (1994a) formally introduced the concept. Scholars have debated the existence of audience costs, the efficacy of their "lock-in" effect, and the causal mechanisms explaining their operation and consequences. However, prior research focused exclusively on the "right-hand side" of the crisis bargaining tree: the consequences of issuing a threat but later backing down. If the consistency mechanism explains a reduction in public support for failing to follow through on a threat, then it should also be operative for failing to honor a public commitment not to intervene militarily. The hypothesized mechanism for explaining behavior on the "right side" of the tree generates observable theoretical implications for behavior on the "left side" of the tree. Expanding the analysis to this "left side" provides additional leverage concerning the importance of consistency, international reputation, and the effects of a leader's invoking new information to justify a course of action. Leaders, such as Lincoln and Wilson, have backed into conflicts after making explicit public statements that they would not intervene. Considering the "left side" of the tree can allow us to understand why and how these leaders are or are not punished for inconsistencies between their words and eventual deeds.

We offer three notable findings in this study. First, participants punished the president for backing into military conflicts after publically promising to stay out, as well as for backing out of prior military threats. This suggests that domestic audiences place a high value on

[29]It is possible that most respondents think of reputation (at least with respect to security issues) as honoring positive threats, not negative promises. If this is true, then the lack of concern for reputational consequences might reflect the implicit assumption that reputation is only relevant for positive commitments, that is, a "reputation for resolve." Thus, there may be a gap between the operational definition of reputation as used in our study (honoring both positive threats and negative promises) and the nominal definition of *reputation* shared by many in the discipline (focusing on positive threats). This possibility deserves further research. For one attempt to deal with the difficult analytic issues surrounding the concept of reputation, see Dafoe, Renshon, and Huth (2014).

[30]While morality is not a primary focus in Fearon (1994a) compared to matters of competence and credibility, recent work has demonstrated an important place for moral concerns among the public, even in matters involving the use of force (Kertzer et al. 2014; Tomz and Weeks 2013). Based on suggestions from an anonymous reviewer, we ran a later round of the survey that included an additional item asking respondents how strongly they felt the president

handled the situation in the "morally right way." Similar to the main results for presidential approval, responses were lower for inconsistent behavior, though the effect was weaker for "Backing in" compared to "Backing down." Likewise, justifying the change in course by citing new information mitigated moral condemnations for failing to follow through on either type of commitment. While only suggestive, the findings suggest a promising additional rationale for why the public may punish leaders' foreign policy behavior and offer a fruitful avenue for further research.

consistency, regardless of the type of commitment. Second, although costs exist for inconsistency across both types of commitments, the president was punished more for failing to implement a public threat to intervene militarily than for failing to honor a public commitment not to intervene militarily. That is, inconsistency costs are greater for backing out than for backing in. Third, and relatedly, the president's pointing to new information to justify his or her change in policy generated a greater reduction of inconsistency costs when backing out of a military commitment than when backing into a military conflict after promising to refrain from the use of force.

In the process of exploring an important but neglected implication of the key causal mechanism driving audience costs theory—the consistency between a leader's commitments and actions—we have raised additional questions, particularly about the theoretical implications of negative commitments. Are the costs of violating negative commitments equal in magnitude to those for violating positive commitments? Are there differences between threats and promises? Do the signals sent by promises to stay out of a conflict help prevent crisis escalation by minimizing the extent to which targeted states misjudge the willingness of their allies to intervene (Benson 2012)? Another question raised by promises to stay out of a conflict—and one amendable to experimental investigation—concerns the form those promises take, and how publics and adversaries respond to subtle variations in the framing of promises. Statements like "we will not intervene," "we will not intervene at this time," and "under no circumstances will we ever intervene" each carry different implications, as does making no promise whatsoever, but instead, just refraining from making any statement at all.

Our study also suggests other areas for future research regarding audience costs. It would be useful to analyze non-American audiences, to see if peoples in states with different levels of military and economic power, political systems and cultures, historical experiences, and saliences of foreign policy issues respond in similar ways. Another significant step forward in audience costs theory would be to examine the beliefs of political leaders and external adversaries, whose behavior is central to audience costs theory. When deciding to issue a deterrent threat, do leaders consider the public's likely reaction—both to the initial threat and to subsequent actions, depending on the adversary's response? Do leaders (and adversaries) take their opponent's potential for domestic audience costs into account when making or responding to threats? These questions are not as amenable to experimental investigation as is the behavior of domestic publics, given problems of external validity.[31] Observational approaches, despite their limitations, may be a further useful way of exploring audience costs theory's many claims and implications for understanding the decisions of political leaders, both domestic and foreign.

References

Basler, Roy P. (ed). 1953. *Collected Works of Abraham Lincoln (CWL)*. Vol 4. Abraham Lincoln Association. New Brunswick, NJ: Rutgers University Press.

Benson, Brett V. 2012. *Constructing International Security: Alliances, Deterrence, and Moral Hazard*. Cambridge, UK: Cambridge University Press.

Berinsky, Adam J. 2009. *In Time of War: Understanding American Public Opinion from World War II to Iraq*. Chicago, IL: University of Chicago Press.

Berinsky, Adam, Gregory Huber, and Gabriel Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 351–68.

Bernauer, Thomas, Robert Gampfer, and Aya Kachi. 2014. "European Unilateralism and Involuntary Burden-sharing in Global Climate Politics: A Public Opinion Perspective from the Other Side." *European Union Politics* 15(1): 132–51.

Broz, J. Lawrence. 2002. "Political System Transparency and Monetary Commitment Regimes." *International Organization* 56(4): 861–87.

Brutger, Ryan. 2014. "When to Compromise? The Influence of Compromise, Proposal Power, and Partisanship on Domestic Audience Costs." Unpublished Manuscript.

Chaudoin, Stephen. 2014. "Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions." *International Organization* 68(1): 235–56.

Cheney, John, Thomas Harford, and Leonard Solomon. 1972. "The Effects of Communicating Threats and Promises upon the Bargaining Process." *Journal of Conflict Resolution* 16(1): 99–107.

Dafoe, Allan, Jonathan Renshon, and Paul Huth. 2014. "Reputation and Status as Motives for War." *Annual Review of Political Science* 17: 371–93.

Davies, Graeme, and Robert Johns. 2013. "Audience Costs among the British Public: The Impact of Escalation, Crisis Type, and Prime Ministerial Rhetoric." *International Studies Quarterly* 57(4): 725–37.

Davis, James W., Jr. 2000. *Threats and Promises: The Pursuit of International Influence*. Baltimore, MD: Johns Hopkins University Press.

Devlin, Patrick. 1974. *Too Proud to Fight: Woodrow Wilson's Neutrality*. New York: Oxford University Press.

---

[31]For a promising example using elite samples, see Hafner-Burton et al. (2014).

Fearon, James D. 1994a. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88(3): 577–92.

———. 1994b. "Signaling versus the Balance of Power and Interests." *Journal of Conflict Resolution* 38(2): 236–69.

———. 1997. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *Journal of Conflict Resolution* 41(1): 68–90.

Goemans, Heins. 2008. "Which Way Out? The Manner and Consequences of Losing Office." *Journal of Conflict Resolution* 52(6): 771–94.

Guisinger, Alexandra, and Alastair Smith. 2002. "Honest Threats: The Interaction of Reputation and Political Institutions in International Crises." *Journal of Conflict Resolution* 46(2): 175–200.

Hafner-Burton, Emilie M., Brad L. LeVeck, David G. Victor, and James H. Fowler. 2014. "Decision Maker Preferences for International Legal Cooperation." *International Organization* 68(4): 845–76.

Heilman, Madeline E. 1974. Threats and Promises: Reputational Consequences and Transfer of Credibility. *Journal of Experimental Social Psychology* 10 (4): 310–24.

Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations." *International Organization* 57(3): 587–616.

Jervis, Robert. 1970. *The Logic of Images in International Relations*. Princeton, NJ: Princeton University Press.

———. 1991. "Domino Beliefs and Strategic Behavior." In *Dominoes and Bandwagons*, eds. Robert Jervis and Jack Snyder. New York: Oxford University Press, 20–50.

Kertzer, Joshua D., and Ryan Brutger. Forthcoming. "Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory." *American Journal of Political Science*.

Levendusky, Matthew, and Michael Horowitz. 2012. "When Backing Down is the Right Decision: Partisanship, New Information, and Audience Costs." *Journal of Politics* 74(2): 323–38.

Levy, Jack S. 2012. "Coercive Threats, Audience Costs, and Case Studies." *Security Studies* 21(3): 383–90.

Martin, Lisa. 2000. *Democratic Commitments: Legislatures and International Cooperation*. Pr inceton, N.J.: Princeton University Press.

Rubin, Jeffrey Z., and Roy J. Lewicki. 1973. A Three-Factor Experimental Analysis of Promises and Threats. *Journal of Applied Social Psychology* 3 (3): 240–57.

Poast, Paul. Forthcoming. "Lincoln's Gamble: Fear of Intervention and the Onset of the American Civil War" Security Studies.

Sartori, Anne E. 2005. *Deterrence by Diplomacy*. Princeton, NJ: Princeton University Press.

Schelling, Thomas C. 1960. *Strategy of Conflict.* Cambridge, MA: Harvard University Press.

———. 1966. *Arms and Influence.* New Haven: Yale University Press.

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

Simmons, Beth. 2010. "Treaty Compliance and Violation." *Annual Review of Political Science* 13: 273–96.

Slantchev, Branislav. 2006. "Politicians, the Media, and Domestic Audience Costs." *International Studies Quarterly* 50(2): 445–77.

Smith, Alastair. 1998. "International Crises and Domestic Politics." *American Political Science Review* 92(3): 623–38.

Snyder, Jack, and Erica Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *American Political Science Review* 105(3): 437–56.

Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61(4): 821–40.

Tomz, Michael, and Jessica L. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(3): 849–65.

Trachtenberg, Marc. 2012. "Audience Costs: An Historical Analysis." *Security Studies* 21(1): 3–42.

Trager, Robert, and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party." *American Journal of Political Science* 55(3): 526–45.

Uzonyi, Gary, Mark Souva, and Sona N. Golder. 2012. "Domestic Institutions and Credible Signals." *International Studies Quarterly* 56(4): 765–76.

Weeks, Jessica L. 2008. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization* 62(1): 35–64.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix A: Instrument for the Main mTurk Survey

Appendix B: Summary Statistics and Representativeness of Sample for the Main mTurk Survey

Appendix C: Robustness Checks from Analysis of Main mTurk Survey

Appendix D: Follow-up Survey Experiment

Appendix E: Empirical Results from Follow-up Survey Experiment