Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory

Joshua D. Kertzer  Harvard University
Ryan Brutger  Princeton University

How are leaders able to credibly signal their intentions in foreign policy crises, despite considerable incentives to bluff? According to a considerable volume of research in international relations, one technique governments have at their disposal is the creation of domestic audience costs: by publicly threatening to use force in order to change an opponent’s behavior, leaders can effectively tie their own hands since their domestic constituencies punish governments that say one thing and do another (Fearon 1994; Smith 1998; Baum 2004; Slantchev 2006; Weeks 2008; Tarar and Leventoglu 2009). In the nearly twenty years since Fearon’s seminal article, a large volume of scholarship has explored the scope and examined the logic of audience costs. Experiments have been central to this endeavor (e.g., Tomz 2007; Hoffman et al. 2009; Trager and Vavreck 2011; Levendusky and Horowitz 2012; Davies and Johns 2013; Brutger 2014; Chaudoin 2014), both because of their ability to avoid the selection effects that make audience costs difficult to study with observational data—if audience costs are real, strategic leaders should avoid making the empty threats that would incur the public’s wrath in the first place (Schultz 2001; Tomz 2007)—and, because at least in democracies, audience cost models rely on a particular set of assumptions about what the public wants (Weeks 2008).

This body of research has told us at least two things. First, audience costs are about inconsistency: whether because of instrumental concerns about the country’s reputation or normative concerns about national honor, publics dislike leaders who say one thing and do another (Fearon 1994; Tomz 2007). Second, audiences are homogenous: The kinds of individual-level characteristics that routinely make a large difference in the study of public opinion about foreign policy seem to play relatively little role in moderating the impacts of treatments in audience cost experiments (Tomz 2007; Levendusky and Horowitz 2012; Davies and Johns 2013). Existing work has thus
assumed that audience costs abide by an essentially unitary logic, and that leaders face an essentially unitary audience.

We argue here that both these assumptions are wrong. Because of how audience costs are defined—the decrease in public support for a leader who threatened the use of force and did not follow through, compared to a leader who simply stayed out—audiences can punish the leader for saying one thing and doing another, but also for threatening the use of force in the first place, a sunk cost we refer to here as a belligerence cost. Failing to take the possibility of belligerence costs into account means we risk misreading what these classic experiments are telling us, and misunderstanding the heterogeneous nature of public responses to crisis bargaining. We present a novel experimental design where we disentangle these two competing logics of why audiences punish, and turn to a series of dispositional characteristics from political psychology and the public opinion about foreign policy literature to bring the audience into audience cost theory. Our results suggest three main points. First, there is no unitary logic of audience costs, as audiences punish leaders both for being inconsistent and for threatening force, such that traditional audience cost experiments may be overestimating how much the public cares about inconsistency. Second, different types of people demand audience costs for very different reasons: hawks and conservatives, for example, punish the President for being inconsistent, while doves and liberals punish the President predominantly for threatening to use force. Both inconsistency costs and belligerence costs can be broadly thought of as types of audience costs in that the audience negatively reacts to the leader’s handling of a foreign policy crisis, but only inconsistency costs follow the traditional audience cost logic put forth by Fearon (1994), whereas belligerence costs invoke a different logic altogether. Third, the presence of these two distinct signaling mechanisms has important implications for crisis bargaining, from explaining the rarity of public threats to showing how the credibility of threats depends not just on the content of the President’s actions, but also on the character of the President’s constituency.

**Audience Costs, Revisited**

Audience cost experiments test a proposition derived from classic crisis bargaining games, whose structure is illustrated in the game tree in Figure 1.¹ The game has two players, represented here as players 1 and 2, who we can think of as the government of a state and its foreign challenger, although the game can also be used to model threats in negotiations more generally. Player 1 decides whether to threaten the use of force against player 2; if player 1 chooses not to threaten force, the game ends with player 1 deciding to stay out. If, on the other hand, player 1 issues a threat, player 2 must decide whether or not to concede. If player 2 does not concede, player 1 then must decide whether to follow through on the threat and engage player 2, or to back down and not engage. Audience cost models are chiefly concerned with the difference between staying out and not engaging; as long as domestic audiences prefer the former to the latter, leaders have to be concerned about domestic political repercussions when they renege on their public commitments. The presence of these costs renders public threats informative rather than merely cheap talk: if leaders expect their domestic constituency will punish them if they issue a threat and back down, they will only issue public threats if they actually intend to follow through; foreign challengers, aware of the existence of audience costs, thus will concede before the crisis can escalate further.

¹See, for example Bueno de Mesquita and Lalman (1992); Fearon (1994); Guisinger and Smith (2002).
leaders are, in fact, rewarded for backing down (Brown and Marcum 2011; Snyder and Borghard 2011; Levendusky and Horowitz 2012; Trachtenberg 2012; Mercer 2012). For our purposes, however, two considerations are especially worth noting here. First, although scholars use multiple mechanisms to explain the existence of audience costs, from concerns about national honor or reputation, to signals of leaders’ quality or competence, embedded in both accounts is the notion that audiences have preferences for consistency: that the public punishes leaders when their words do not match their deeds (Fearon 1994; Smith 1998; Chaudoin 2014). Indeed, one of the major charges levied against audience cost theory by its critics is that audience cost models erroneously assume the public cares about consistency rather than the policy outcome itself (Clare 2007; Snyder and Borghard 2011; Downes and Sechser 2012).

Second, although a large body of literature has investigated the scope of audience cost theory by focusing on the nature of the domestic political institutions in which the audience is embedded—for example, whether audience costs vary with regime type, opposition criticism, or the leader’s rhetoric (Slantchev 2006; Weeks 2008; Baum 2004; Levendusky and Horowitz 2012; Trager and Vavreck 2011; Davies and Johns 2013; Brutger 2014), there has been relatively little attention paid to the relevant characteristics of audiences themselves. This asymmetry stems partially from the prevailing tendency of IR scholars to focus predominantly on situational features rather than dispositional characteristics, partially because formal approaches to audience cost theory tend to assume a homogenous audience for purposes of analytic tractability, and also because those experimental studies that have looked for heterogeneous treatment effects have tended to find that the logic of audience costs is impervious to the types of individual differences that routinely play an important role in public opinion about foreign policy more generally. The exception to this rule has been work on political identification and engagement, although the findings have been mixed (Tomz 2007; Levendusky and Horowitz 2012; Trager and Vavreck 2011; Davies and Johns 2013). In short, the experimental literature on audience costs has tended to interpret the large audience cost treatment effects as a sign that publics care about consistency, and the general lack of heterogeneous treatment effects as a sign that the logic of audience costs manifests itself across different subsets of the population.

Audience Costs as a Double-barreled Treatment Effect

We believe political scientists should exercise caution before drawing these conclusions, because there are more to audience costs than meets the eye. One of the fundamental tasks in designing and analyzing experiments is establishing the relevant counterfactual, selecting the control group to be compared to the treatment (Morgan and Winship 2007). In audience cost theory, as Figure 1 illustrates, both the treatment and the control are clearly specified: the treatment is the President threatening force and failing to follow through, while the control is the President deciding to stay out of the conflict in the first place. In this respect, however, the treatment and the control differ from one another in two different ways: both in terms of whether the President says one thing and does another, but also whether the President threatens to use force in the first place.

Since experimentalists and crisis bargaining scholars often speak subtly different languages, we make this point in two different ways in Figure 2. The left-hand side of Figure 2 models leaders’ decisions in audience cost theory as experimentalists would, with a $2 \times 2$, in which the vertical axis refers to leaders’ initial decision about whether to threaten the use of force or not (which we refer to here as $\alpha$), and the horizontal axis depicts the subsequent decision about whether to back down or follow through on the initial decision (which we refer to here as $\beta$). The top-left cell (with expected value $\mu_{11}$) represents the control group from audience cost experiments, in which a leader decides to stay out, while the bottom-right cell (with expected value $\mu_{22}$) represents the classic audience cost treatment, in which a leader threatens the use of force, and fails to follow through. Since the audience cost treatment effect is defined as $\mu_{22} - \mu_{11}$—the average difference in responses across the main diagonal in the $2 \times 2$—the treatment and control differ from one another in both $\beta$ and $\alpha$, both the leaders’ deeds, and the leaders’ words: not just whether the leader follows through, but also whether she threatens to use force in the first place. In this sense, the audience cost treatment effect is “double-barreled.”

To be clear, our claim is not that audience cost experiments are unfaithful to audience costs, but rather that...
because the model the experiments are testing involves comparing outcomes at two different levels of the game tree, both the experiments and the model they are based on inevitably employ an “indissoluble double stimulus” (Converse and Presser 1986, 13). Indeed, we express the same argument using a revised crisis bargaining model in the right-hand side of Figure 2, which we turn to later. More important for our purposes is that this double-barreled treatment effect has implications both for how we understand the logic of audience costs, and how we test whether the same logic of audience costs manifests itself across different segments of the population. First, there are a variety of reasons why members of the public might disapprove of the President threatening force, ranging from doves who believe that military options only make problems worse (Holsti 1979), to Jeffersonians who would rather the United States focus on its own problems (Mead 2002), to realists who see the President getting involved in a situation where the national interest is not at stake as a sign of a lack of competence (Kertzer and McGraw 2012). If the total audience cost treatment effect represents both the impact of inconsistency and the impact of threatening force, then there is no longer a unitary logic of audience costs. Different audiences can punish leaders for very different rationales, and we may wish to explore how much of the total audience cost comes from concerns about inconsistency versus concerns about the use of force. Second, if audiences punish leaders for different reasons, then simply looking at whether the size of the audience cost treatment effect varies across different segments of the population obscures the possibility that the magnitude of the treatment effect might remain the same even if the mechanism behind it varies dramatically.

To disentangle these logics, we adopt a mid-way point between the traditional audience cost experiment and a fully-crossed 2 x 2 factorial, illustrated by the left-hand panel of Figure 2: we supplement the classic two-condition audience cost experiment with a third experimental condition in which the leader threatens the use of force and follows through—the bottom-left cell (with expected value $\mu_{11}$ in the 2 x 2). This simple modification is important because, as Figure 2 shows, this “Engage” experimental condition is a middle ground between the “Stay Out” condition and the “Not Engage” condition: it differs from the “Not Engage” condition in terms of whether the leader is consistent (the presence of the threat of force is held constant), and differs from the “Stay Out” condition in terms of whether the leader threatens the use of force (consistency is held constant). The absence of the fourth cell means we cannot analyze the experimental results as if the experimental design were a traditional 2 x 2 factorial, since the two factors are not fully crossed.

There are a number of alternative approaches that could be used to disentangle these logics, including the use of open-ended responses, which we discuss along with a number of alternative designs in Appendix §3. In an innovative exploration of crisis bargaining models, Trager and Vavreck (2011) also measure public approval at the “Engage” node of the model, but because their experiment was designed to answer a different set of questions, they vary military outcomes rather than holding them constant, which precludes us from being able to decompose the audience cost treatment into inconsistency costs and belligerence costs in their design, for reasons we explain below.
However, since the “Engage” experimental condition ($\mu_{21}$) is a middle ground between the “Stay Out” condition ($\mu_{11}$) and the “Not Engage” condition ($\mu_{22}$), the total audience cost treatment effect $\mu_{22} - \mu_{11}$ can be re-expressed as the sum of two separate treatment effects:

$$\mu_{22} - \mu_{11} = (\mu_{21} - \mu_{11}) + (\mu_{22} - \mu_{21}) \quad (1)$$

To make this analytical move, we need to ensure the experiment holds outcomes constant across conditions: after all, we cannot capture how much the audience dislikes the threat of force if staying out results in splendid isolation while going in leads to a costly quagmire. More formally, we need to isolate the treatment effects from the audience’s perception of payoffs. In the classic crisis bargaining model displayed in Figure 1, actors make decisions based on the expected payoffs at the terminal nodes of the game tree. The initiating state is assumed to be fighting over some policy or good with value $v$ (usually normalized to a value of 1), and in the event it makes a threat and follows through, its expected payoff is $w_{11}$, a war payoff that incorporates the state’s belief about the probability of winning the good of value $v$ and some expected cost paid for engaging in military action. In observed crises, it would be nearly impossible to disentangle audiences’ approval for the value of the policy outcome from the strategy used to achieve it. The advantage of experiments, however, is that we can ensure participants are presented with an identical outcome (worth $v$) regardless of whether the President stays out, engages, or does not engage, such that changes in public approval can be traced entirely to the leader’s strategy. We can therefore revise the payoffs from the classic crisis bargaining model in Figure 1 in two ways: first, denoting these identical outcomes across the nodes of the game tree, and second, decomposing audience costs by adding a new payoff, $-f$, reflecting the possibility that audiences punish leaders for threatening force rather than just for being inconsistent.

Six points are important to note about this payoff structure, depicted in the game tree in the right-hand side of Figure 2. First, incorporating the $-f$ term renders transparent the double-barreled nature of audience cost models, since the audience cost is now $-f - a$: the sum of two different treatment effects. Second, although the notation is different, the game tree in the right-hand side of Figure 2 corresponds exactly to the experimental design in the left-hand side. Comparing approval at the Stay Out and Engage nodes of the tree isolates $-f$, the belligerence cost—the simple effect of threatening force ($\mu_{21} - \mu_{11}$ from the $2 \times 2$), while comparing approval between the Engage and Not Engage nodes produces $-a$, the inconsistency cost—the simple effect of inconsistency ($\mu_{22} - \mu_{21}$). Third, it is worth re-emphasizing that $-f$ and $-a$ are simple effects rather than main effects; we refer interested readers to Appendix §4 for a discussion of why the former are the quantities of interest in audience cost models, rather than the latter.

Fourth, if threatening force has no impact on respondents’ evaluations of the President (that is, $-f = 0$), then the conventional wisdom about audience costs will be correct, in that we should interpret the audience cost treatment effect as being driven by concerns about inconsistency, and all of the various mechanisms (national honor, leaders’ competence, etc.) it implicates, since if $-f = 0$, then $-f - a = -a$. If, on the other hand, $-f \neq 0$, at least some of the audience cost treatment effect is driven by concerns about the threat of force, and there is no longer a unitary logic of audience costs. This point can be understood both as a narrower methodological claim, and as a broader substantive one. On the one hand, it is clear that incorporating a double-barreled design means that classical audience cost experiments offer a systematically biased estimate of how much people care about inconsistency. This point is subtly different from critiques raised by Snyder and Borghard (2011), Chaudoin (2014) and others: Our concern is not so much that people care less about inconsistency than audience cost theory assumes, but that because of their double-barreled design, these traditional experiments cannot tell us how much people care about inconsistency at all. On the other hand, however, these classical experiments are inheriting the assumptions of the crisis bargaining model they are based on; the potential for belligerence costs thus raises interesting implications for crisis bargaining scholars more generally, with important ramifications for how leaders send signals in international crises.

Fifth, audience costs have traditionally been understood as a way for leaders to tie their hands: unlike sunk-cost signals that are immediately costly for leaders to carry out, but should not affect their future behavior, hands-tying incentivizes future behavior by producing costs that are only paid if the leader does not follow through (Fearon 1997). If audiences primarily punish leaders for being inconsistent, then public threats should be informative signals for crisis bargaining and function in the manner consistent with audience cost theory, since leaders will pay the cost only if they back down on the threat. If, however, audiences punish leaders for making threats, then public escalation generates sunk costs, since the cost is paid when the threat is made, thereby invoking an alternative signaling mechanism and undermining the prevailing understanding of the signaling value of audience costs.

Finally, if audience costs have two distinct logics, then we should be able to estimate these quantities of interest across different types of respondents (hawks versus doves,
nationalists who see their country as superior to others, versus cosmopolitans who identify with the broader global community, etc.) and observe the extent to which the logic of audience costs vary. If we find that some audiences punish leaders for making threats and others punish inconsistency, then the signaling value of public escalation will be even murkier as different audiences impose costs for different reasons. This means signaling in crises is contingent on both the President’s strategy and the relevant audience. To effectively interpret the signaling value of public escalation, a foreign adversary would have to evaluate the relative importance of different audiences to the leader, which would make public escalation a noisier signal in international crises than it is often considered.

**Bringing the Audience Back In**

One of the advantages of decomposing audience costs into concerns about inconsistency versus concerns about the threat of force is that it opens up the possibility of integrating audience cost theory with a large body of research in political psychology and public opinion about foreign policy that explores the dispositional underpinnings of foreign policy attitudes. As noted above, existing audience cost experiments have largely shied away from investigating how demands for audience costs vary with dispositional characteristics of audiences, but those investigations that have taken place have tended to search for dispositional effects by testing whether the size of the audience cost treatment effect varies across subgroups (e.g., Tomz 2007; Davies and Johns 2013). However, if audience cost treatment effects actually contain two different quantities of interest, the logic of audience costs might vary dramatically between hawks and doves, say, even if the magnitude of the audience cost does not significantly differ. There is no shortage of dispositional characteristics we could investigate, but for theoretical reasons we focus on four characteristics here, each one corresponding to a slightly different logic.

First is militant assertiveness. Although, as Weeks (2008, 43) reminds us, “the ‘audience cost’ does not arise because domestic audiences disagree with their leaders’ policy,” if audience costs actually involve two different costs—one reflecting inconsistency, and the other reflecting the threat of force—then understanding variation in audience costs requires us to model substantive policy preferences as well (Snyder and Borghard 2011; Levy 2012; Chaudoin 2014). A large volume of research in public opinion on foreign policy tells us that the use of force is an issue about which people's attitudes vary systematically (Holsti 1979; Chittick, Billingsley, and Travis 1995; Murray 1996; Herrmann, Tetlock, and Visser 1999; Rathbun 2007; Gelpi, Feaver, and Reifler 2009). Beliefs about the desirability and effectiveness of the use of force form a major part of all of the classic models in the American foreign policy attitudes literature, even if this construct takes a number of different names: Wittkopf (1990) (and later, Holsti and Rosenau 1990) refer to “militant internationalism”; Hurwitz and Peffley (1987), to “militarism”; and Herrmann, Tetlock, and Visser (1999), to “militant assertiveness.” Similar constructs have been employed in analyses of the structure of foreign policy attitudes in countries outside the United States as well (Hurwitz, Peffley, and Seligson 1993; Bjereld and Eken gren 1999; Reifler, Scotto, and Clarke 2011). Audiences should care far less about inconsistency if they disapproved of the policy the leader is backing down from in the first place: thus, since hawks and doves differ in their beliefs about whether governments should use military means to achieve their foreign policy ends, we expect that individuals who are low in militant assertiveness will demand audience costs primarily to punish the President for threatening to use force, but will be less concerned about inconsistency, while those who are high in militant assertiveness should support the threat of force, and thus, demand audience costs for failing to use it.

Second is international trust. According to one of the classic interpretations of audience cost models, audiences dislike inconsistency because it sullies the country’s reputation for resolve and weakens its future bargaining position (Guisinger and Smith 2002; Tomz 2007). In this context, like in many rationalist models, reputation and trust are inextricably linked since institutional environments that incentivize the building of the former facilitate the emergence of the latter (e.g., Milgrom, North, and Weingast 1990; Kydd 2005). Yet as Rathbun (2009, 2011) argues, this type of “strategic trust” based on incentive structures is at odds with much of the psychological and sociological literature on trust, which also understands trust as moralistic or fiduciary, based on beliefs about the character of general or specific others (Cook 2001; Uslaner 2002; Hoffman 2002; Hardin 2006). Whereas strategic trust is situationally derived, moralistic trust is dispositionally varying: some people are generally more trusting than others, endowed with a sunnier view of human nature that “allows one to go beyond incentives” (Mercer 2010, 6), be more prone to cooperation, and less concerned about exploitation (Kuhlman and Wimberley 1976; Wrightsman 1991; Kanagaretnam et al. 2009). In this sense, it should not be surprising that variation in

---

5 In Uslaner’s 2002 framework, moralized trust can be particularized (“I trust you”) or generalized (“I trust”); many of the kinds of trust
trust has been found to predict attitudes toward a wide range of foreign policy issues, from the use of force to support for free trade (Brewer and Steenbergen 2002; Chanley 2002; Brewer 2004; Binning 2007; Kaltenthaler and Miller 2013). Individuals who display less international trust, who are more cynical about the prospects of exploitation by others and thus should be most concerned about maintaining reputations for resolve, should demand audience costs primarily to punish the President for displaying inconsistency, and should hardly be bothered by the President’s threat of force. In contrast, individuals with high levels of international trust, who are more sanguine about the motivations of other countries and should thus be less preoccupied with maintaining reputations for resolve, should be more likely to punish presidents for making threats than for inconsistency.

Third is nationalism. As Tomz (2007, 835) notes, many reputation-based accounts of audience cost models offer what one might call an instrumental or consequentialist (March and Olsen 1998) understanding of audience costs, in which audiences punish inconsistency because of the negative consequences they believe result from it. However, we can also think of audience cost mechanisms as following not an instrumental concern about reputation or credibility, but a moral or normative concern about national honor. Thus, another key audience characteristic relevant for our purposes is what Chittick, Billingsley, and Travis (1995) call “identity,” Rathbun (2007) calls “community,” and what social psychologists would call “ingroup identification”: the extent to which individuals are nationalists with deep identifications with their nation (and thus, more willing to use force on its behalf), or cosmopolitans who identify with the international community as a whole. Building off of work on social identity theory—which distinguishes between ingroup favoritism and outgroup derogation (Brewer 1999)—IR scholars who have studied ingroup identification tend to distinguish between national attachment versus national chauvinism, the latter of which examines individuals’ tendencies to not only identify with their nation, but also view it as superior to others (Herrmann, Isernia, and Segatti 2009). National chauvinists, who have little qualms about using force on their ingroup’s behalf, should punish the President not for threatening the use of force, but for violating the national honor by being inconsistent. In contrast, individuals who are low in national chauvinism, who are less likely to believe in the inherent inferiority of other countries, and thus, less willing to resort to force, should be relatively less sensitive to inconsistency and more to the threat of force.

Fourth is political ideology—the broader belief system that structures how people think about politics, typically represented as a continuum with liberals on the left, and conservatives on the right (Converse 1964; Jost et al. 2003; Feldman 2003). Although characteristics like international trust and militant assertiveness are helpful in offering clearly-specified microfoundations for public opinion, holistic constructs like ideology are also useful because they encapsulate these more specific traits within a broader framework. A plethora of psychological research emphasizes the motivational underpinnings of political ideology, positing that individuals are attracted to liberal or conservative belief systems to fulfill discrete psychological needs (Jost et al. 2003, 2007). Conservatives are higher not just in needs for order, structure, and cognitive closure (Webster and Kruglanski 1994) that should make them less tolerant of inconsistency, but also in authoritarianism, social dominance, and system justification (Altemeyer 1998; Sidanius and Pratto 2001; Duriez and Van Hiel 2002; Jost, Banaji, and Nosek 2004), which should make them more acceptant of the use of force. They are more likely to subscribe to “competitive jungle” and “dangerous world” beliefs (Duckitt et al. 2002), that should enhance their concern about reputations for resolve, while serving as license for the use of force. These existential and epistemic motives cluster together, but point in the same direction: we should expect conservative audiences to punish leaders for inconsistency, while liberal audiences will be relatively more concerned about the use of force.

Two additional considerations are worth keeping in mind. First, each of our theoretical expectations outlined above suggest that at high and low levels of these dispositional characteristics, audience costs may follow a logic of substitutable rather than conjunctive causation (Bräumoeller 2003): some segments of the audience will
DECOMPOSING AUDIENCE COSTS

demand inconsistency costs and others belligerence costs, rather than punishing for both reasons simultaneously. Second, needless to say, there are a far larger number of audience characteristics we could potentially examine than the four traits we discuss above. To the extent that we find evidence for heterogeneous treatment effects across these four traits, it is reasonable to assume that we would find further evidence of heterogeneity across other audience characteristics as well.

Method

To test our theoretical framework, we fielded an online survey experiment in the spring of 2014 on a national American sample of 942 registered voters recruited by Survey Sampling International (SSI). The study includes an audience cost experiment (described in detail below) as well as a dispositional questionnaire measuring the series of individual-level characteristics outlined above, allowing us to examine the extent to which different types of members of the public demand audience costs for different types of logics. Since the purpose of the study is to challenge the conventional interpretation of audience costs, the structure of the audience cost experiment borrows heavily from the classic audience cost experiments described above. In the scenario, respondents were first presented with an introduction that read:

The following questions are about U.S. relations with other countries around the world. You will read about a situation our country has faced many times in the past and will probably face again. Different leaders have handled the situation in different ways. We will describe one approach U.S. leaders have taken, and ask whether you approve or disapprove.

Respondents were then presented with information about a hypothetical international crisis, which, like most audience cost experiments, evokes the classic “repel an invader” experiment first introduced in IR by Herrmann, Tetlock, and Visser (1999), in which a foreign country sends its military to take over part of a neighboring country. Following existing audience costs experiments, we deliberately chose not to identify the foreign countries in the scenario so as to avoid introducing country-specific confounding, and to speak more directly to the existing literature. After participants read that “A country sent its military to take over a territorial region in a neighboring country,” respondents were randomly assigned to one of three different experimental conditions, reflecting the three different cells in Figure 2. In the “Stay Out” condition, participants were told that the President announced that “the United States would stay out of the conflict,” whereupon the attacking country continued to invade. In the “Not Engage” condition, participants were told that the President announced that “if the attacking country continued to invade, the U.S. military would immediately engage and attempt to push out the attacking country”; the attacking country continued to invade, but the President did not engage.

Thus, as in all audience cost experiments, the “Stay Out” condition and the “Not Engage” condition vary both (i) whether the President threatens the use of force, and (ii) whether the President says one thing but does another. We therefore introduce a third experimental condition as a middle ground between these two treatments: in the “Engage” condition, the President announces that “if the attacking country continued to invade, the U.S. military would immediately engage and attempt to push out the attacking country”; the attacking country continued to invade, and the President ordered the U.S. military to engage. The treatment thus differs from each of its counterparts in an important manner: it differs from the “Not Engage” treatment in terms of whether the President displays inconsistent behavior, and differs from the “Stay Out” condition in terms of whether the President’s initial commitment is to stay out or to use force. To control for policy outcomes, across all three treatments the final result for the United States is the same: the conflict ends with the hostile state gaining 20% of the contested territory regardless of whether the President goes in or stays out, and because the war payoff generally includes an additional cost of fighting, we specify in the “Engage” condition that the United States experienced zero casualties in the conflict. Doing so not only minimizes the perceived cost of war, but also serves as a more conservative test; as we show in Appendix §5, replicating the experiment without specifying the absence of casualties in the “Engage” condition produces larger estimates of the cost of threatening force. In all three conditions, we measure our dependent variable of interest by asking participants the extent to which they approved or disapproved of the way the President handled the situation, which yielded a seven point approval scale.

We also include a questionnaire measuring a series of dispositional characteristics borrowed from the political psychology and public opinion about foreign policy literatures: we measure participants’ militant assertiveness using items borrowed from Herrmann,
Tetlock, and Visser (1999), national chauvinism using items from Herrmann, Isernia, and Segatti (2009), international trust from Brewer (2004), and political ideology using the standard measure from the American National Election Studies (ANES). We also measure a variety of standard demographic characteristics, including gender and education.

**Results**

We present our results in two phases. First, we simply look at our treatment effects, decomposing audience costs into concerns about inconsistency versus concerns about threatening force. Second, we use the individual differences measured in our dispositional questionnaire to show how different kinds of respondents weigh these two rationales differently, such that audience costs can be demanded for very different reasons by different audiences.

**Average Treatment Effects**

As an initial examination of the results, we estimate the audience cost treatment effect $\mu_{22} - \mu_{11}$, the average difference in support for the President when the President threatens force and backs down, compared to the President choosing to stay out of the crisis. As is the case in all audience cost experiments, the audience cost treatment is negative and statistically significant: approval for the President is 0.45 points lower (95% bootstrapped CI: -1.71, -0.98) when the President backs down on a threat than when she follows through on staying out in the first place. However, as previously discussed, this treatment effect is double-barreled: it contains not only the costs produced from backing down, but also those incurred from threatening to get involved in the first place. When we decompose this treatment effect into its constituent parts, we can see that our participants indeed punish the leader for being inconsistent, in that $\mu_{22} - \mu_{21}$—the average difference in support for the President when the President threatens force and backs down versus threatens force and follows through—is negative and statistically significant: on average, approval for the President is 0.90 points lower (95% bootstrapped CI: -0.90, -0.02) when the President follows through on a threat to use force than on a pledge to stay out.\(^{10}\)

Three points are worth noting here. First, note that the total audience cost treatment effect is a sum of inconsistency costs and belligerence costs; this is not simply a coincidence of the data, but axiomatically true, since by definition, $(\mu_{21} - \mu_{11}) + (\mu_{22} - \mu_{21}) = \mu_{22} - \mu_{11}$. Second, as Gartzke and Lupu (2012) point out, audience costs are about a specific mechanism, rather than just an effect. Yet these results show us that audience cost experiments are measuring more than just concerns about inconsistency: 67% of the audience cost treatment effect in our sample comes from an inconsistency cost, but 33% comes from a belligerence cost. Third, although classical audience cost studies have tended to treat the public as homogenous, the public opinion literature has tended to find considerable heterogeneity both in terms of what people think about foreign affairs and how they think about it. As we see below, the presence of these two different forms of punishment offers one potential reason why it has been surprisingly difficult to find evidence of heterogeneous treatment effects.

**Heterogenous Logics of Audience Costs**

The conventional way of investigating heterogeneity with audience costs would be to estimate a series of statistical models in which we interact the audience cost treatment with a number of dispositional characteristics to see whether certain types of people demand larger audience costs than others (e.g., Tomz 2007; Levendusky and Horowitz 2012; Davies and Johns 2013). Our interest here, however, is somewhat different, since we wish to disentangle the relative importance of each of these two logics. In lieu of complex statistical models, we keep the analyses simple by estimating a series of subgroup analyses, where we split the sample in two (separating hawks from doves, for example), and decompose the audience cost in each subgroup, estimating what fraction of the audience cost effect comes from inconsistency costs ($\frac{\mu_{22} - \mu_{21}}{\mu_{22} - \mu_{11}}$), versus belligerence costs ($\frac{\mu_{21} - \mu_{11}}{\mu_{22} - \mu_{11}}$).\(^{11}\) In Appendix §6, we replicate these analyses in a regression context that controls

\(^{10}\)We can also examine belligerence and inconsistency costs by measuring changes in the percentage of the audience that disapproves of the President’s handling of the situation (Tomz 2007, 835). For the full sample, inconsistency increases disapproval by 21 percentage points, whereas belligerence increases disapproval by 8 percentage points. For additional results in percentage point form, see Appendix §7.

\(^{11}\)For all subgroup analyses, we define low and high levels of each dispositional characteristic using the interquartile range, thus comparing individuals in the bottom and top 25%, but this method obviously generalizes to the specification of other cutpoints as well.
for all of the dispositional variables simultaneously, and show the substantive results remain the same.

We present the results both numerically in Table 1 and visually using density plots in Figure 3. The first three columns of Table 1 display the three treatment effects (audience costs, belligerence costs, and inconsistency costs), while the two right-hand columns decompose the audience cost effect, showing our two main quantities of interest: the fraction of the audience cost effect that comes from belligerence costs versus from inconsistency costs. These two quantities of interest are displayed visually in Figure 3, in which the fraction of audience costs stemming from inconsistency is illustrated in white, and the fraction of audience costs stemming from concerns about the use of force is displayed in dark gray, the probability distributions are derived from 2,000 bootstraps to provide estimates of uncertainty. Thus, the first row of the table and panel a of Figure 3 show that as discussed above, 67% of the audience cost effect in our full sample comes from inconsistency costs, while 33% comes from belligerence costs. While our results replicate existing studies that find significant audience costs, it is also clear that we risk seriously misunderstanding the logic underpinning why audiences punish leaders in domestic crises if we attribute the presence of audience costs solely to the public’s punishment of inconsistency. The question is whether this pattern replicates when we start disaggregating the audience.

The first dispositional characteristic we look at is militant assertiveness, which illustrates how policy preferences drive audience costs. When we compare those participants whose militant assertiveness scores in the dispositional survey placed them in the bottom 25% (panel b) versus those in the top 25% (panel c), we can see that doves and hawks demand audience costs for fundamentally different reasons. We find no evidence that respondents who are low in militant assertiveness are turned off by inconsistency—indeed, the inconsistency treatment effect is not statistically significant from zero (p < 0.49)—but are heavily concerned about the leader threatening force (p < 0.00), which is responsible for the entire cost they impose. In contrast, although those respondents who are high in militant assertiveness also demand an audience cost, it is entirely driven by the leader being inconsistent. Two points are noteworthy here. First, although the idea that doves and hawks should have different preferences about foreign policy might seem unsurprising, this finding challenges all previous experimental work on audience costs we are aware of, which has found hawks and doves to be relatively homogenous in how they demand audience costs, and thus, assumes there to be very little variation in how much people care about leaders who display inconsistent behavior. Second, simply looking at the magnitude of the audience cost effect between doves and hawks would be misleading: Doves in our sample actually demand larger audience costs than hawks do, but this effect is a consequence of hawks rewarding the President for threatening force; in our sample, focusing solely on the magnitude of audience costs thus not only overestimates how much doves care about inconsistency, but also

<table>
<thead>
<tr>
<th>Audience Cost</th>
<th>Belligerence Cost</th>
<th>Inconsistency Cost</th>
<th>Belligerence Cost Fraction</th>
<th>Inconsistency Cost Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) (\mu_{22} - \mu_{11})</td>
<td>(2) (\mu_{21} - \mu_{11})</td>
<td>(3) (\mu_{22} - \mu_{21})</td>
<td>(4) (\frac{\mu_{21} - \mu_{11}}{\mu_{22} - \mu_{11}})</td>
<td>(5) (\frac{\mu_{22} - \mu_{21}}{\mu_{22} - \mu_{11}})</td>
</tr>
<tr>
<td>(-a - f)</td>
<td>(-f)</td>
<td>(-a)</td>
<td>(-\frac{a}{a-f})</td>
<td>(-\frac{f}{a-f})</td>
</tr>
</tbody>
</table>

Full Sample | \(-1.355\) | \(-0.447\) | \(-0.906\) | \(0.326\) | \(0.674\) |
Low Militant | \(-1.770\) | \(-1.754\) | \(-0.017\) | \(1.016\) | \(0.016\) |
High Militant | \(-1.284\) | \(0.194\) | \(-1.478\) | \(-0.225\) | \(1.225\) |
Low Int Trust | \(-1.304\) | \(-0.135\) | \(-1.169\) | \(0.090\) | \(0.910\) |
High Int Trust | \(-1.462\) | \(-1.548\) | \(0.086\) | \(1.173\) | \(-0.173\) |
Low Nat Chauv | \(-1.122\) | \(-0.522\) | \(-0.599\) | \(0.459\) | \(0.541\) |
High Nat Chauv | \(-1.554\) | \(-0.535\) | \(-1.019\) | \(0.337\) | \(0.663\) |
Liberals | \(-1.058\) | \(-0.870\) | \(-0.188\) | \(0.882\) | \(0.118\) |
Conservatives | \(-1.516\) | \(-0.386\) | \(-1.130\) | \(0.224\) | \(0.776\) |

Note: Columns (1) through (3) depict average treatment effects calculated from 2,000 bootstraps, shown with both experimental and crisis bargaining notation; note that the threaten force cost (2) and inconsistency cost (3) add up to the total audience cost (1). Columns (4) through (5) calculate the fraction of the total audience cost stemming from threatening force (4), versus from inconsistency (5). Note that these two fractions add up to 1. For more on these quantities of interest, please refer to Figure 2.
FIGURE 3 The Multiple Logics of Audience Costs

Note. For both the full sample (panel a) and the series of subsamples depicted in panels b-h, Figure 3 decomposes the audience cost treatment effect $\mu_{22} - \mu_{11}$ into two separate treatment effects: concerns about threatening force ($\mu_{21} - \mu_{11}$) in dark grey and concerns about inconsistency ($\mu_{22} - \mu_{21}$) in white, each one expressed as a fraction of the total audience cost treatment effect, and illustrated as density distributions. The x-axes are bounded at (-0.4,1.4) for ease of interpretability. These density plots show that audience costs display multiple logics: hawks (panel c) and doves (panel b), for example, demand audience costs for very different reasons.

underestimates how much hawks care about inconsistency by 15%.

The second characteristic we turn to is international trust, which, as noted above, offers one way of testing the reputational mechanism posited to drive audience costs. Since individuals who are lower in international trust are more cynical about their country’s exploitation by others, and thus, have the most reason to be preoccupied with maintaining reputations for resolve, we would expect them to demand audience costs primarily to punish the President for displaying inconsistency, and thereby, harming the country’s bargaining reputation. In contrast, individuals with higher levels of international trust, guided by sunnier views of the motives of other countries, should be less concerned about inconsistency and more about the threat of force. Sure enough, 91% of the audience cost for respondents low in international trust is driven by concerns about inconsistency (panel d of Figure 3), while respondents high in international trust (panel e) demand audience costs entirely because of the threat of force. On the one hand, then, these findings offer evidence in favor of reputational understandings of audience costs, but on the other hand, it also reminds us that not all audiences are concerned about reputational costs.

Third, we turn to national chauvinism, which ties into accounts of audience costs that stress violations of the “national honor.” We articulated above why nationalists, who are the most likely to embrace notions of national honor, should be more likely to punish the President for inconsistency, while cosmopolitans, less likely to see the United States as superior to other nations, should care less about inconsistency, and more about the threat of force. Panels f and g of Figure 3 offer some support for this hypothesis: nationalists who see the United States as far superior to other nations strongly punish the President for being inconsistent (panel g), while relatively “cosmopolitan” respondents low in national chauvinism (panel f) display relatively weak effects, split between the two logics ($p < 0.10$ for inconsistency costs, $p < 0.14$ for belligerence costs). However, unlike with militant assertiveness and international trust, the difference in the fraction of audience costs driven by inconsistency across the two subgroups is not statistically significant ($p < 0.381$). At least as proxied by national chauvinism, then, we find less evidence for national honor-based explanations of
audience costs than for the other mechanisms tested here.

Finally, panels h and i of Figure 3 look at political ideology, disaggregating audience costs among self-reported liberals versus conservatives. As expected, audience costs among liberals are driven primarily by belligerence costs, while a full 78% of conservatives’ audience cost comes from inconsistency. Once again, then, we find signs of heterogeneous logics of audience costs, showing that classical audience cost experiments are painting a somewhat distorted picture of the dynamics of public opinion in foreign policy crises: both liberals and conservatives demand audience costs, but for very different reasons.

Implications

These dual forms of audience costs provide a series of substantive insights into the value of public threats as signals in international crises, four of which we briefly touch on here.

First, audience costs include a sunk cost. Until now, audience cost scholars have focused on signaling through hands-tying, in that threats were understood to generate costs that were only paid if the leader chose not to follow through. In contrast, we find that leaders also pay a sunk belligerence cost when they make threats: in a population composed of both doves and hawks, for example, the President loses the support of the doves as soon as she issues the threat of force, and does not regain their approval regardless of whether she follows through or not. Having already made a threat, her level of support now depends on the approval of hawks, whose adamant aversion to inconsistency means that she can either follow through on her commitment, or back down and alienate both segments of population. The leader is thus faced with an initial decision of whether to pay a belligerence cost when the threat is made, while also generating additional potential costs that will only be paid at a later date, depending on her decision to follow through or not.

Second, threats can screen: even when an audience does not punish their leader for failing to follow through on a threat, the public threat itself generates a domestic belligerence cost that directly affects the leader’s approval (and payoff in the crisis bargaining model), which can thus be used as a costly signaling mechanism, enabling foreign adversaries to learn about the threatening state’s level of resolve. Since a leader would only initiate a threat when the expected value of the prize is greater than the expected cost of the threat,12 initiating a threat acts as a credible screening device. Only leaders who place a sufficiently high value on the contested prize would be willing to publicly threaten, which means foreign leaders can update their beliefs when public threats are observed.13 In this sense, although IR scholars have generally understood diplomatic communication to be of limited value in the absence of traditional audience costs (Fearon 1995; Guisinger and Smith 2002; Smith 1998), our findings offer additional support in favor of Sartori’s (2002) and Ramsay’s (2011) arguments that talk in IR is less cheap than is often claimed. This point also illustrates an important difference from other work critiquing audience cost theory: Snyder and Borghard (2011), for example, suggest that empty threats are less costly than audience cost theory assumes, while our findings about belligerence costs suggest that threats can be costly regardless of their emptiness.

Third, leaders have an incentive to pursue secret diplomacy. An expanding literature on the value of secret negotiations highlights the ability of closed-door communications to play an important role in bargaining (Kurizaki 2007; Ramsay 2011; Trager 2010; Yarhi-Milo 2013). Most notably, Yarhi-Milo (2013) outlines the informative role of secret communication when the revelation of the communication may have domestic costs. While Yarhi-Milo focuses on assurances, the same logic holds in the presence of belligerence costs. The threatening leader can avoid the sunk cost by conveying the threat in secret, yet the threat remains credible given that there is a risk the cost will be paid if the threat is later revealed.14 This suggests that there is no such thing as costless bluffing when it comes to public threats, and leaders have an incentive to exercise restraint when considering whether to initiate a threat under the eyes of a watchful public. As Sechser (2011) notes, public coercive threats are surprisingly scarce, and the presence of belligerence costs offers one potential explanation.

Fourth, the credibility of public threats depends on who the leader’s relevant audience is, and the results from the “heterogeneous logics” section illustrate a public sufficiently variegated that interpreting signals is all the more complicated (Jervis 1976). Not only do leaders’ key constituencies change over time thanks to the ebb and flow

12Any strategy where the threatening state pays a sunk cost greater than the value of the prize is strictly dominated (Fearon 1997).
13Baum (2004) also notes that leaders should evaluate the potential reward when making threats, but the key decision for the leader in his formulation is weighing the reward versus the risk of military defeat, which invokes a different calculation since military defeat is a potential cost contingent on state power, as opposed to a sunk domestic cost.
14It remains a task for future research to determine whether a leader may talk her way out of this type of punishment by providing a justification for her strategy, as Levendusky and Horowitz (2012) have argued can be done with traditional audience costs.
of domestic politics, but so too do the distributions of each of the audience characteristics we explored here, which tend to fluctuate in response to events at home and abroad (MacKuen, Erikson, and Stimson 1989; Keeler 2007; Kertzer 2013). Thus, one would expect that as the composition of the public shifts, so too would the balance of considerations driving audience costs. Moreover, a host of psychological biases from the foreign policy decision-making literature suggests foreign leaders are thus likely to miscalculate the credibility of public threats in the face of mosaic audiences: belief perseverance and embedded images mean decision-makers will be slow to update their beliefs in response to these changes (Herrmann and Fischerkeller 1995), while the outgroup homogeneity effect makes it harder for decision-makers to pick up on divisions in rival foreign publics (Judd and Park 1988). The twin dynamics of belligerence and inconsistency costs thus offers further grist for Mercer’s (2012) concern about how cognitive limitations can contort the effects of signals, and invites additional skepticism about audience costs’ efficacy.

Conclusion

Given what Baum and Potter (2014) call the “cottage industry” of scholarship on audience cost theory, audience cost experiments have become one of the more popular experimental traditions in International Relations, despite facing a flurry of critiques. Some critics note that Presidents can escape punishment for inconsistency by pointing to new information, managing elite criticism, or by making concessions in private (Levendusky and Horowitz 2012; Saunders 2015; Brown and Marcum 2011). Other skeptics express concern that the logic of audience costs seems to better manifest itself in the lab rather than in crucial historical cases (Snyder and Borghard 2011; Trachtenberg 2012). Another group still suggests that citizens have preferences over policy choices and outcomes that may trump preferences about consistency (Snyder and Borghard 2011; Downes and Sechser 2012; Chaudoin 2014). We argue here that even when we set these external critiques of audience cost theory aside, we need to be careful about the inferences we draw about the logic of audience costs, since the control and treatment groups differ in multiple ways, such that participants can punish both in response to inconsistency and for the leader threatening the use of force.

We presented the results here from a novel experimental design where we disentangle these two competing logics of why audiences punish. Our results show that there is no unitary logic of audience costs, and that some audiences impose inconsistency costs while others impose belligerence costs. Bringing the audience back into our analysis gives us a richer understanding of the dynamics of public responses to crisis bargaining, and thus, helps narrow the gulf between audience cost theory and the study of public opinion. The results presented here suggest that audience costs are not governed by a fundamentally different logic from that which governs public attitudes toward foreign policy issues more generally, in that the dispositional characteristics used by political psychologists and public opinion scholars to explain variations in foreign policy attitudes also explain differences in how people demand audience costs. These dual logics also raise important implications for crisis bargaining, explaining the rarity of public threats, and showing how their signaling value is contingent on the leader’s relevant constituency.

References


Brown, Jonathan N., and Anthony S. Marcum. 2011. “Avoiding Audience Costs: Domestic Political Accountability and


**Supporting Information**

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

Appendix §1: Dispositional measures
Appendix §2: Sampling methodology and sample characteristics
Appendix §3: Alternative experimental designs
Appendix §4: Simple effects versus main effects
Appendix §5: The impact of specifying casualty levels on the cost of threatening force
Appendix §6: Regression analysis
Appendix §7: Alternative measure of approval