The Rationality of Audience Costs: Does Engaging the National Honor Allow For Credible Signals?

Shawn Ling Ramirez

Prepared for MPSA 2012

Abstract

Many theoretical and empirical works rely upon expectations derived from audience costs theories in which leaders think twice before issuing threats given the domestic punishment, or audience costs, they will face if they back down from these threats. All of these works assume that audience costs exist and act in the manner we expect, and assume that the underlying mechanism for why these audience costs are imposed is due to the audience's concern for national honor or reputation. I model a domestic public that is concerned for its international reputation – the audience receives the value of the updated beliefs about its own country's type as a component of its payoff in addition to the crisis outcome – in a model of crisis bargaining. My results show that the traditional effect of audience costs exists only in one region of this game. I show that that audience costs are necessary, but not sufficient conditions for credible signals of resolve, and further, that audience costs can also be imposed to intentionally mislead the enemy about type – audience costs may be deceptive. By showing that the effects of audience costs are non-monotonic, my work expands upon the role of international reputation in generating the potential effects of audience costs, but also raises new questions about the limits of empirical tests and the theoretical literature thus far.

Keywords: audience costs

Contents

1 Introduction

Since Fearon's (1994) seminal work demonstrated that audience costs – the threat of domestic punishment that a leader will suffer if he backs down from his threats – are a mechanism by which leaders could learn in international crises in spite of their incentives to misrepresent and private information, numerous theoretical and empirical works have shows how audience costs can contribute to our understanding of crises, war, peace and the effect of domestic political institutions on international relations. Yet, the theory is not without its critics. Many questions regarding why the audience imposes costs, what makes audience costs rational, and what is the mechanism that generates audience costs are often left out of the theoretical and empirical equation. The reasoning that is given (which goes back to Fearon's 1994 article) is that the audience has some innate concern for their own country's international reputation or honor. But this brings up a further question: why doesn't the leader share this concern? What creates this rift between the preferences of the audience and the leader? My work attempts to answer this question by theoretically modeling an audience that is concerned for its international reputation, and asking under what condition will it strategically impose audience costs?

I adapt a model of crisis bargaining in which there are two countries in crisis, and one of those countries has a leader and a domestic audience. The audience receives a payoff for the crisis outcome as well as for its international reputation earned during the crisis; the beliefs about the audience's country's type, namely, whether it is strong or weak feed directly into the audience's preferences. After Nature draws the type of the home country, the audience chooses the probability that it will impose audience costs – the extent of audience costs that the leader will suffer if he backs down from the threat that initiated the crisis – in light of the potential ramifications that audience costs will have on their own international reputation. If the audience cost literature has sufficiently specified the underlying mechanism – that a concern for reputation allows for audience costs which allow for credible signaling – then my

results should be in line with expectations from audience cost theories.

Unsurprisingly, I find that credible signaling is possible in only part of the overall parameter space describing leader behaviors (Lemma 1). More surprisingly, I find that audience costs are a necessary but not sufficient condition for credible signals of resolve (Result 3), and only the lack of audience costs is sufficient to demonstrate incredible resolve (Result 1). Further, I specify the three conditions under which audience costs will be imposed (Result 2).

As a whole, my results show that audience costs do not necessarily lend themselves to credible signals of resolve. Further, that sometimes high audience costs may be intended to mislead the enemy about a country's true type. When the country is weak, and the war will be lost, but common knowledge suggests that the country is most likely strong, the audience prefers to impose audience costs and encourage all types to escalate (rather than back down) to benefit when common knowledge encourages the enemy to concede and enhances the country's reputation.

To proceed, section 2 discusses the literature. Section 3 describes the model. Section 4 evaluates the model and discusses the results. Section 5 concludes. All figures are in the appendix.

2 The Debate about Audience Costs

Since Fearon's seminal work in 1994 that established the theoretical basis by which state leaders are able to demonstrate their resolve credibly through the mechanism of audience costs, an extensive and probing literature has evolved. Numerous theoretical works enhance our understanding of the effect of audience costs at various stages in an international crisis and across different types of political regimes. Numerous empirical works look for and find

evidence of the effects of audience costs to varying degrees. Yet, a deeper question persists about the logic of audience costs. Why do audience impose these costs and why should audience need to impose these costs? First I highlight some of the recent findings in the literature and the debate around audience costs. Then I argue that the answers to these persistent questions have not been satisfactorily addressed.

According to Fearon (1994), audience costs are costs that the leader will suffer if he backs down from conflict that increase as the crisis escalates: "a leader who chooses to back down is (or would be) perceived as having suffered a greater 'diplomatic humiliation' the more he had escalated the crisis" (p. 580). This fits the intuitions that leaders suffer from diplomatic humiliation for escalating crises from which they back down, and that leaders enjoy the fruits of a diplomatic triumph for escalating crises in which they stand firm – as their enemy backed down.

The implications of audience costs also fit with intuitions about the abilities for democracies to resolve the security dilemma, shedding some light on a possible mechanism for the democratic peace. In Fearon's model, leaders who face the threat of audience costs will only stand firm insofar as they can tolerate the audience costs they would suffer – and therefore, leaders are able to credibly signal their resolve through escalation (and the rising threat of audience costs). From this, Fearon notes the theoretical implication that leaders of democracies, who may face a greater threat of domestic costs, may be better able to send credible signals of resolve than leaders of autocracies who are often shielded from their domestic audiences.

Additional theoretical work has found audience costs useful in accounting for behaviors at other stages in the crisis bargaining game. Kurizaki (2007) shows that audience costs can make threats credible even when bargaining in secret: the public nature of audience

costs need not hold in order for audience costs to be effective. Beardsley (2010) finds that potential domestic audience costs increases the chances that mediation occurs. Ramirez (2011, unpublished manuscript) answers the question of how this occurs by showing that leaders who face potential audience costs are advantaged when bargaining for peace through a third-party mediation process: the threat of audience costs enables leaders to secure better peaceful settlements, or if bargaining breaks down, wars against weaker enemies.[1]

Further work has demonstrated that there are subtleties as to what makes audience costs a democratic phenomenon, or whether audience costs are a democratic phenomenon at all. Schultz (2000) shows that one way in which audience costs in democracies might operate is through the domestic political opposition, who provides a signal of endorsing or dissenting against the leader's international threat. Ramsay (2004) shows that when a domestic political opposition is motivated by potential electoral outcomes and potential international crisis outcomes, the domestic opposition will form a credible and truthful signal to both international and domestic audiences. Slantchev (2006) argues that audience costs are higher in democracies only under specific conditions, namely, when if the freedom of the press is sufficiently protected. Weeks (2008) finds that autocracies are also able to generate audience costs as long as political elites are able to coordinate in a way that would be observable to the foreign enemy.

However, there is much debate about whether Fearon's argument allows for testable implications. Schultz (2001) was perhaps the first to note that since leaders choose their strategies (to stand firm) when facing the potential threat of audience costs, the audience are never suffered – these audience costs are only suffered off the equilibrium path, and therefore there is great empirical difficulty in testing whether audience costs existed and influenced the leader's behavior. However, the first to attempt to test Fearon's implications statistically,

[1]Ramirez, Shawn L. "Diplomatic Options in the Shadow of An Audience: The Benefits of Private Mediation." *unpublished manuscript.*

and to find results in his favor, was Partell and Palmer (1999): they find that although national capabilities is a strong indicator of crisis outcomes, the influence of domestic political audiences (operationalized by the democratic nature of the state) does effect crisis outcomes in a positive way. In accord with Fearon's theory, Schultz (2000) finds that enemies are more likely to back down when threats are issued by leaders of democracies. Gelpi and Griesdorf (2001) test numerous mechanisms that support the democratic peace phenomenon on international crises in the 1900s. They find that democratic political structures lead to positive crisis outcomes, more strongly than norms, trade, or shared cultures – again, lending support to the mechanism that the ability to generate audience costs allows for credible signals of resolve. Experimental work in a novel approach by Tomz (2007) demonstrates that audience costs – costs that the leader would pay for making threats and then backing down – exist across a variety of conditions.

Yet, there is also evidence against the theory that audience costs allow for credible signals of resolve. In an examination of case studies during the Cold War, Snyder and Borghard (2011, unpublished manuscript) argue that audience costs rarely if ever play a critical role in a leader's decision to issue threats, and that the public does not hold the leader accountable for failing to act in accordance with threats – rather the public cares about "the overall substantive consequences of the leader's policy." Downes and Sescher (2010, unpublished manuscript) question the validity of statistical tests that examine the utility of audience costs in making threats credible (and successful), since these statistical tests rely on data that include very few if any threats.

However, there is a fundamental question that remains unanswered about audience costs. Both the theoretical and empirical works that have built from Fearon (1994) have proceeded to evaluate the extent to which the implications of audience cost theories are true, but have failed to uncover the reason why audiences impose these costs in the first place. In other

words, all work thus far has taken the stance that: assuming that audience costs exist, then audience costs should have the following effects. But why do audiences impose or generate these audience costs? In Fearon (1994), the underlying mechanism was that leaders chose to "engage the national honor." The domestic audience preferred that the leader make only those threats that he would not back down from because otherwise the audience would suffer international honor or reputational costs. In order to bring the leader's actions in line with the preferences of the audience, the audience imposes costs. This is supported in the experimental results by Tomz (2007), where he finds that the reason that audience costs are imposed is because "citizens care about the international reputation of the country or leader."

Thus far, no theoretical work to my knowledge asks the question of whether the domestic audience will impose costs if it is concerned with international reputation or national honor. All models that invoke audience cost theories assume that in the larger game (preceding the model at hand), the audience cares about its honor or reputation, and that given this concern the audience will impose audience costs against a leader who backs down from his threats. As a result our observations of the "coercive effect" of audience costs should fall in line with expectations derived from Fearon (1994), and the numerous theoretical results that have expanded the effects of audience costs beyond this.[2] In this model, I assume that the audience has concern for its international reputation – the posterior belief about their own country's type after the outcome of the crisis (reputation if the leader backed down, stood firm, escalated, won, etc.) – and I ask, given this concern, under what circumstances will audiences impose audience costs?

[2]Slantchev's unpublished manuscript "Audience Costs Theory and Its Audiences" also raises this complaint, among others, and calls this the expected effects of audience costs the "coercive effect."

3 The Model

Two countries, 1 and 2 (or home and enemy) are involved in a crisis over a prize w that the enemy country has, and the home country wants.[3] Country 1 consists of a leader and a domestic audience, and country 2 is a unitary actor. The crisis is modeled in a game of one-sided incomplete information. At the start of the game, Nature draws country 1's type which gives the probability that country 1 will win in a war against country 2; $p \in \{p_L, p_H\}$, where $0 \leq p_L < p_H \leq 1$, and the probability that $p = p_H$ is given by $q \sim U[0,1]$. Both the leader and the audience observe their country's true type, which is their private information, while country 2 remains uncertain.

Next, to investigate the role of audience costs and whether they are rationally imposed, the audience has the choice of whether or not they would like to impose audience costs.[4] That is, after Nature draws the home country's type, the audience chooses $\theta \in [0,1]$ which gives the probability that the audience will impose audience costs. Audience costs are represented by $a > 0$, which are the costs that the leader will suffer if he backs down from the crisis.

The leader then decides whether to back down, attack the other country, or to escalate the crisis. If the leader backs down, then the leader pays audience costs with probability θ. Country 2 keeps the prize, giving it a "no loss" payoff of zero. The audience does not receive the prize, but it does receive a payoff for the international reputation its country earned as a result of the crisis. This international reputation is the updated belief (the posterior belief) about the type of the audience's home country; the audience receives the value of the posterior as its payoff. If the leader attacks, then the two countries fight a war. The war is

[3] I assume that $w > 0$ and that $w \leq 1$. This makes war possible, but not automatically so.

[4] This allows me to investigate when audience costs are rational, given that the audience knows the truth about its own country's chances of winning in war. An interesting follow-up question can then be asked: when audiences don't know their country's true type, and leaders are more knowledgeable than the public, will leaders want to manipulate the beliefs of the audience in order to ratchet up their own potential audience costs?

modeled as a costly lottery where each country pays a cost for fighting, $c_1 > 0$ and $c_2 > 0$. If country 1 wins the war with probability p, then country 1 receives the prize w and country 2 loses the prize (receives a payoff of $-w$). If country 2 wins the war, then country 2 keeps the prize (receives a payoff of zero) and country 1 receives a payoff of zero. The leader receives country 1's war payoff. The audience also receives their country's war payoff as well as the international reputation payoff (the posterior belief about the type of country 1 given that the leader attacked).

Alternatively, the leader can escalate (continue) the crisis. The enemy updates its beliefs about the type of country 1, and chooses to back down or to attack. If the enemy backs down, then the enemy concedes the prize. The leader earns w, and the enemy receives $-w$. The audience receives its country's payoff of w plus the international reputation payoff. If the enemy attacks, then the two states fight the war with similar payoffs as above.[5]

Therefore, at every outcome, the audience receives a payoff for the international reputation earned from the actions their leader has taken during the crisis. Given this concern for international reputation, how and when do audiences impose audience costs?

4 Evaluation and Results

Familiar Territory: Enemy and Leader Strategies

To see what the enemy does at the last node, let μ represent the enemy's belief that $p = p_H$ after the leader escalates. The enemy's best response is to attack if

$$\mu \leq \frac{w - c_2}{w},$$

[5]There are no first strike advantages, nor penalties for costly delay.

and back down otherwise.

The leader chooses between backing down, escalation, and attacking by choosing the strategy that provides the highest utility, $\max\{\theta(-a), pw - c_1, kw + (1-k)(pw - c_1)\}$, where k represents the probability that the enemy will back down after escalation. Note that escalate weakly dominates attack for all $k > 0$ since any possibility of concessions makes escalation a better option:[6]

$$kw + (1-k)(pw - c_1) > pw - c_1.$$

Therefore the leader only needs to choose between backing down and escalation. The leader prefers to back down if:

$$\begin{aligned} \theta(-a) &\geq kw + (1-k)(pw - c_1) \\ \theta &\leq \frac{kw + (1-k)(pw - c_1)}{-a} \equiv \theta^*(p). \end{aligned}$$

Intuitively, the leader prefers to back down if the audience is not punishing very often, i.e., with probability θ that is low enough, $\theta \leq \theta^*(p)$. This value depends on the type that is drawn, however, here I will refer to $\theta^*(p)$ as θ^*.

Since θ^* is decreasing in p, this threshold is lower for high types than for low types. Let $\bar{\theta}^*$ represent the threshold for $p = p_H$, such that if $\theta < \bar{\theta}^*$ then the leader backs down, and if $\theta > \bar{\theta}^*$ then the leader escalates when $p = p_H$. Let $\underline{\theta}^*$ represent the threshold for $p = p_L$, such that if $\theta < \underline{\theta}^*$ then the leader backs down, and if $\theta > \underline{\theta}^*$ then the leader escalates when $p = p_L$.

Expanded Territory: Leader Strategies Traditionally Placed Aside

[6] This relies on there being no escalation costs that might occur with delay or mobilization.

The best responses of the leader divide the parameter space θ into three regions, where $0 \leq \bar{\theta}^* \leq \underline{\theta}^* \leq 1$. For $\theta \in [0, \bar{\theta}^*)$, the leader will back down regardless of type (region 1). For $\theta \in [\bar{\theta}^*, \underline{\theta}^*)$, the leader will escalate if the type is high, $p = p_H$, but back down if type is low, $p = p_L$ (region 2). Note that region 2 is the range in which the role of audience costs in crisis bargaining is normally examined: in region 2 audience costs separate stronger types from weaker types and allow stronger types to signal their resolve credibly. For $\theta \in [\underline{\theta}^*, 1]$, the leader will escalate regardless of type (region 3). These regions are described in Lemma 1.

Lemma 1.

Depending on the probability that audience costs are imposed, θ, where $\bar{\theta}^* = \frac{kw+(1-k)(p_H w - c_1)}{-a}$ and $\underline{\theta}^* = \frac{kw+(1-k)(p_L w - c_1)}{-a}$, the best response of the leader is to:

- *back down when $\theta \in [0, \bar{\theta}^*)$*
- *escalate if $p = p_H$ and back down if $p = p_L$ when $\theta \in [\bar{\theta}^*, \underline{\theta}^*)$*
- *escalate when $\theta \in [\underline{\theta}^*, 1]$.*

What remains to be determined are the optimal strategies for the audience and to see whether they form an equilibrium. In region 1, since a leader will back down regardless of type, country 1 receives none of the prize and there is no new information about the type of the country from the leader's actions. The audience's payoff is q.

In region 2, audience costs separate the types, therefore the actions of the leader are informative about the type that was drawn and the posterior beliefs are one or zero, depending on whether the leader escalated or backed down, respectively. When $p = p_L$, the leader will back down, the country receives nothing, and the international reputation of the country is that it is a low type – the audience receives a payoff of zero. When $p = p_H$, the leader will escalate, the international reputation of the country is that it is a high type, and the enemy's

updated belief $\mu = 1$. Given these beliefs, the enemy will back down since $\mu = 1 > 1 - \frac{c_2}{w}$. The audience receives a payoff of $w + 1$.

In region 3, both types escalate, and the posterior beliefs $\mu = q$. When a low type is more common, $q < 1 - \frac{c_2}{w}$, the enemy will attack, and the audience will receive a payoff of $pw - c_1 + q$. When a high type is more common, $q > 1 - \frac{c_2}{w}$, the enemy will back down, and the audience will receive a payoff of $w + q$.

Optimal Audience Costs For The Weak Given Expanded Territory

To summarize the possibilities given the best responses of the leader and the enemy, if a low type is drawn, $p = p_L$, then the audience can receive a payoff of q in region 1, zero in region 2, and either $pw - c_1 + q$ when $q < 1 - \frac{c_2}{w}$ or $w + q$ when $q > 1 - \frac{c_2}{w}$ in region 3. First, note that if a low type is drawn, and if the audience knows that its country's reputation will be poor as a result of the conflict (as is assumed by this model), then it never makes sense for the audience to impose audience costs in the manner in which scholars traditionally examine audience costs (in region 2).

Instead, as long as the expected value of war is sufficiently high, $pw - c_1 > 0$, the audience's preferred strategy is to impose audience costs with a high probability, setting θ in region 3, so that the leader will escalate regardless of type. This strategy for the audience keeps the enemy uncertain about the strength of the home county. In doing so, by raising audience costs to encourage the leader to escalate (no matter what the type is), the audience can take a shot at obtaining either concessions or the war payoff in addition to the poor international reputation that nature has drawn, $pw - c_1 + q$ or $w + q$.

On the other hand if the expected value of war is sufficiently low, $pw - c_1 \leq 0$, then

the audience only wants to take a chance on war if the enemy is likely to back down upon observing escalation, i.e., when the home country is like to be a high type. When $q > 1 - \frac{c_2}{w}$, the audience imposes audience costs with a high probability so that the leader will escalate and the country can obtain $w + q$. Notice that in this case it is as if the audience capitalizes on the high expectation of its country's strength that encourages concessions by the enemy, even though in reality it has been drawn an unlucky hand, $p = p_L$. When $q \leq 1 - \frac{c_2}{w}$ and the expectation of the country's strength is low, the audience sets a low probability of imposing audience costs, θ is in region 1, so that the leader will back down regardless of type, and the audience earns only the low reputation that it already had.

Optimal Audience Costs For The Strong Given Expanded Territory

If a high type is drawn, the audience can expect to receive a payoff of q in region 1, $w + 1$ in region 2, and in region 3, either $pw - c_1 + q$ when $q < 1 - \frac{c_2}{w}$ or $w + q$ when $q > 1 - \frac{c_2}{w}$. Given these possibilities, in all but one case, the audience has a dominant strategy when $p = p_H$ to set θ in region 2.[7] When a high type is drawn, the audience prefers chooses this middle-range of θ in which its country's type is revealed allowing the audience earns a strong reputation and gain concessions. Therefore, the results here are in line with Fearon's assertion that audience costs are sensible ex ante because leaders want the opportunity to reveal their type credibly, however, this model places the extent of audience costs in the hands of the audience – and demonstrates that this level of audience costs would not occur if the type that were drawn were low.

Audience Incentives and Optimal Audience Costs

In total, this means that if the probability of imposing audience costs, θ, is observable

[7] This strategy dominates in all cases except when $q = 1$, in which case region 2 and region 3 provide the same payoff.

to the enemy, then one must ask whether the enemy can detect the type of the country based on the observed value of θ, the extent of audience costs.

The audience sets θ to lie within region 2 if $p = p_H$, but how is region 2 defined? Given the enemy's strategy to back down if escalation occurs, the leader will have no incentive to deviate when the high type is drawn as long as $\theta \geq \bar{\theta}^*$ where $\bar{\theta}^* = \frac{w}{-a}$, which holds for any $\theta \geq 0$.

Interestingly, note that if the low type were drawn and the audience set θ to be in region 2, we would not have an equilibrium since the low type would deviate to escalation: the leader (with a low type) has no profitable deviation if and only if $\theta \leq \underline{\theta}^* = \frac{w}{-a}$, which would require θ to be less than some negative value – not possible. In other words, in the *only* region in which audience costs allow for credible signals, there would not be an equilibrium in which audience costs allow for credible signals of resolve! Lucky for us here, since the audience never chooses region 2 if a low type is drawn, we don't need to worry about this situation in *this* model.

However, for scholars who use audience costs as a means for strong types to signal their resolve credibly, it is important to note that the reason that audience costs allow for credible signals in Fearon's model is because of its *infinite time horizon: the threat of having to back down in the future prevents deviation by the low type.* In this finite model, the low type does not need to back down in the future, since the structure of the game forces the enemy to make that final choice between backing down and standing firm. Therefore, scholars who currently employ audience costs in their models must ask whether or not they truly are examining an infinite time horizon game: do weak or low types genuinely fear backing down in the future (so types separate and resolve can be signaled credibly), or can they corner their enemies into making those choices for them perhaps through international institutions, conflict resolution processes, or outcomes on the battlefield? This leads to the following

conjecture:[8]

Conjecture 1.

> *In any finite crisis bargaining game in which the leader who faces potential audience costs is not forced to make the choice between backing down and standing firm at the final node, i.e., the enemy makes this choice, there is no separating equilibrium – audience costs do not allow for credible signals of resolve.*

What if the low type is drawn? Recall that when $p = p_L$, there are two situations to consider: when the expected value of war is high, and when that expected value is low. First, when the expected value of war is high, the audience imposes costs with a high probability to ensure that the leader will escalate regardless of type. In doing so, the audience keeps the enemy uncertain about the country's type (the low type, $p = p_L$), posterior beliefs are the same as the prior, $\mu = q$, and the enemy will back down if $q > 1 - \frac{c_2}{w}$, and attack otherwise. Therefore, when q is high (the type of the home country is likely to be strong), the value of θ that makes a leader escalate regardless of type is any value of $\theta \geq \underline{\theta}^*$, where $\underline{\theta}^* = \frac{w}{-a}$, which holds for all values of $\theta \geq 0$. When q is low (the home country is likely to be weak), $\theta \geq \underline{\theta}^*$, where $\underline{\theta}^* = \frac{p_L w - c_1}{-a}$ in order to prevent either leader from deviating to back down immediately: note that this also holds for all values of $\theta \geq 0$ since the expected value of war is high, $p_L w - c_1 > 0$.

When the expected value of war is low, $p_L w - c_1 \leq 0$, recall that two further situations arise. Either the home country is likely to be strong, in which case, the audience wants to capitalize on this high expectation and encourage its leader to escalate (regardless of signal), so the enemy will back down. For neither leader to deviate given this strategy profile, the audience must choose any value of $\theta \geq \underline{\theta}^*$, where $\underline{\theta}^* = \frac{w}{-a}$ which holds for all $\theta \geq 0$.

Alternatively, when the expected value of war is low, $p_L w - c_1 \leq 0$, and when the home

[8] I am currently working on proving this result.

country is likely to be weak, $q \leq 1 - \frac{c_2}{w}$, then the audience wants the leader to back down so that none will be the wiser about the true type of the country. For a leader to prefer to back down rather than escalate, the following needs to be true:

$$\theta(-a) \geq kw + (1-k)(pw - c_1).$$

Both leaders will deviate if the enemy backs down, therefore, the beliefs off the equilibrium path must ensure that the enemy attacks. What beliefs will support this? For the enemy to attack after observing escalation, the enemy must believe that $p = p_H$ with probability $\mu < 1 - \frac{c_2}{w}$. Notice that if the audience sets $\theta = 0$, then the leader will prefer to back down rather than to escalate, and further, since this is the only situation in which the audience does not impose audience costs, the enemy knows that the type of the country must be low. The enemy can use the lack of audience costs as a signal of a low type, $\mu = 0$, and attack if escalation is observed. This leads to the following result:

Result 1.

> *The lack of audience costs is a credible signal of the lack of resolve. An audience who imposes audience costs strategically will not impose audience costs only if the expected value of war is low, $p_L w - c_1 \leq 0$ and the home country is likely to be weak, $q \leq 1 - \frac{c_2}{w}$. This allows any leader to back down immediately rather than escalate a crisis into a losing war, and protects the audience's country from earning the reputation of a low type.*

Since that is the final case to explore, we can state the equilibrium here. In stating the equilibrium, the leader strategies can be described as pooling and separating according to the parameter space defined by θ. However, since the audience observes p and chooses the probability of imposing audience costs strategically, the equilibrium paths of play are far more illuminating. Proposition 1 summarizes the equilibrium paths of play for this game,

where the degree of pooling or separating exhibited by the leader strategies can be told according to the region of θ.

Proposition 1. *The equilibrium paths of play are as follows:*

1. *When $p = p_H$, the audience chooses any $\theta \geq 0$, θ is in region 2, the leader escalates, the enemy updates its beliefs, $\mu = 1$, the enemy backs down.*

2. *When $p = p_L$ and $p_L w - c_1 > 0$, the audience chooses any $\theta \geq 0$, θ is in region 3, the leader escalates, the enemy updates its beliefs, $\mu = q$, and the enemy backs down if $q > 1 - \frac{c_2}{w}$ and attacks otherwise.*

3. *When $p = p_L$ and $p_L w - c_1 \leq 0$, then:*

   (a) *if $q > 1 - \frac{c_2}{w}$, the audience chooses any $\theta \geq 0$, θ is in region 3, the leader escalates, the enemy updates its beliefs, $\mu = q$, and the enemy backs down.*

   (b) *if $q \leq 1 - \frac{c_2}{w}$, the audience chooses any $\theta = 0$, θ is in region 1, the leader backs down, the enemy updates its beliefs, $\mu = 0$, and the enemy attacks.*

First, notice that audience costs are imposed with any probability $\theta \geq 0$ in cases 1, 2, and 3(a). In case 1, the type of the country is high. In cases 2 and 3(a), the type of the country is low. Therefore, the presence of the threat of audience costs (and the lack of a threat, since $\theta = 0$ is included in these possible paths of play), do not help to distinguish between times when type is low or high.

Second, audience costs only allow for credible signals of resolve in region 2. However, θ is in region 2 only in case 1. In cases 2 and 3(a), θ is in region 3 where the leader escalates regardless of type, audience costs are imposed, but the audience costs are not credible signals of type. The imposition of audience costs occurs in case 2, because even though the audience knows that it has a low type, the overall value of war is high enough to warrant taking a

chance on war or concessions. Note that resolve in Fearon 1994 was the expected value of war. So this is similar to his result: one can think of these "types" of p_L, when $p_L w - c_1 > 0$, as not the types that are not too low. However, in case 3, this occurs because the audience knows that even though it has drawn a low type, common knowledge about the distribution of q suggests that the type of the home country is likely to be high: by keeping the enemy uncertain and encouraging any type to escalate, the audience bluffs by encouraging escalation that might leave the country with a strong reputation, and possibly even concessions – even though it knows it has drawn a low type and the expected value of war is low. In this case, the audience is not a credible signal of resolve in any possible sense. This forms the following result: audience costs are not sufficient for credible signals of resolve, and further, they can sometimes be meant to mislead.

Result 2.

> *When audience costs are imposed strategically, audience costs will be imposed under three conditions:*
>
> - *if the country is strong,*
> - *if the country is weak but the expected value of war is high enough, or*
> - *if the country is weak and the expected value of war is low, but common knowledge suggests that there is a high probability that the country is strong.*
>
> *This last case suggests that audience costs can be used to mislead the enemy or the international arena.*

Since a credible signal only occurs when there are audience costs imposed, audience costs are a necessary condition for credible signals of resolve. However, since audience costs are imposed in many other cases, that do not involve credible signals of resolve, and sometimes are meant as a misleading signal of resolve, audience costs are not a sufficient condition for

credible signals of resolve.

Result 3.

> *Audience costs are necessary, but not sufficient conditions for leaders to provide a credible signal of resolve in crisis bargaining.*

By controlling when to impose audience costs, the audience has substantial control over the actions that the leader will take and ultimately the audience's payoff. One of the keys to this game is that the payoff is made in terms of the outcome of the crisis, or the pie, as well as the country's international reputation at crisis' end. By allowing the audience to gain an international reputation payoff, at times the audience prefers to take a chance on war regardless of whether they expect the enemy to back down or stand firm. This occurs because when the leader will escalate regardless of type, the "obscured" high international reputation may be better than the country's true type, and the enemy is taking actions based on expecting that high type.

Perhaps the model places too much power in the hands of the audience. One can ask whether it is too strong of an assumption to say that the audience knows the true value of chances that the home country will win in war. Or perhaps it is a lot to assume that the audience will take actions in a coordinated fashion that requires overcoming collective action problems and an enhanced knowledge of the game structure. However, the audience of a country makes estimates of its chances of victory in deciding whether to support its country or its leader in a crisis.

A deeper question is whether and how leaders will want to manipulate the audience's perception of the value of p, the probability of winning in war, in order to manipulate the role of audience costs in a crisis. It may be more plausible to assume that leaders have not only the ability to manipulate the audience's perception of p, but also the audience's knowledge

of the game structure. If in certain circumstances, audience costs can cause an enemy to update its beliefs to $\mu = 1$, while in other cases, the enemy updates its beliefs to $\mu = q$, then leaders may have incentives to manipulate their audience costs in ways not yet explored.

5 Conclusion

In explicitly modeling the assumption that the audience has concern over its country's international reputation, the model provides a contribution in the development of audience cost theories both theoretically and empirically.

First, the model shows that the traditional effect of audience costs used in our scholarly research, by which leaders can credibly signal their resolve, is only true within a portion of the parameter space – only when audience costs are in the "middle range." In the lower range, leaders will back down regardless of type, and in the upper range, leaders will escalate regardless of type. This is not to suggest a non-monotonic effect: the values of the "middle range" depend on numerous other factors in the game in order to stabilize the leader's strategies in equilibrium. Thus in case 1 in Proposition 1, the audience costs are in this "middle range" although empirically we should expect any probability of facing audience costs, $\theta \geq 0$.

Second, I find that given the equilibrium paths of play, there are three instances in which audience costs are strategically imposed. First, audience costs are imposed if the country is strong as in traditional audience cost theories. Second, audience costs are imposed if the country is weak, but the expected value of war is high enough. This case is similar to audience cost theories, in which resolve is still somewhat high for there to be audience costs, however, the extent of audience costs imposed here is the same as the extent of audience costs imposed in the first case – therefore audience costs do not allow for a finer distinction between more and less resolved countries.

Third, and most surprisingly, I find that audience costs will be imposed when the country is weak and the expected value of war is low, but common knowledge suggests that the home country likely to be strong. In this case, the audience wants to mislead the enemy country, and encourage both strong and weak types to escalate using audience costs, so that the enemy cannot guess the true type of the country. Since the home country is likely to be strong, the enemy will concede and the international reputation of the home country will reflect the common knowledge assumption rather than the country's true type.

Further, I find that even though the presence of audience costs does not distinguish any particular case, the lack of audience costs does. The domestic audience will not impose costs if the expected value of war is low and the home country is likely to be weak. Therefore, when there is a lack of audience costs, then this is a credible signal that there is no resolve.

Finally, my results depend on this being a finite game in which the enemy makes the last choice. My results suggest that in any finite game in which leaders face potential audience costs but are not forced to back down or stand firm in the final node of the game, i.e., when the enemy is forced to make this decision, audience costs do not allow for credible signals of resolve. This is because by forcing the enemy to make that move, strong and weak leaders can choose to escalate and obscure their true type. Weak leaders will have incentive to do this as long as they can ensure that the enemy will make the final decision, and the home country's reputation will not be damaged. When might this happen in reality and why is this important? Perhaps on the battlefield, or in conflict resolution processes, or through international institutions, certain countries with high audience costs may be able to force enemies to make the final withdrawal.

6 Appendix

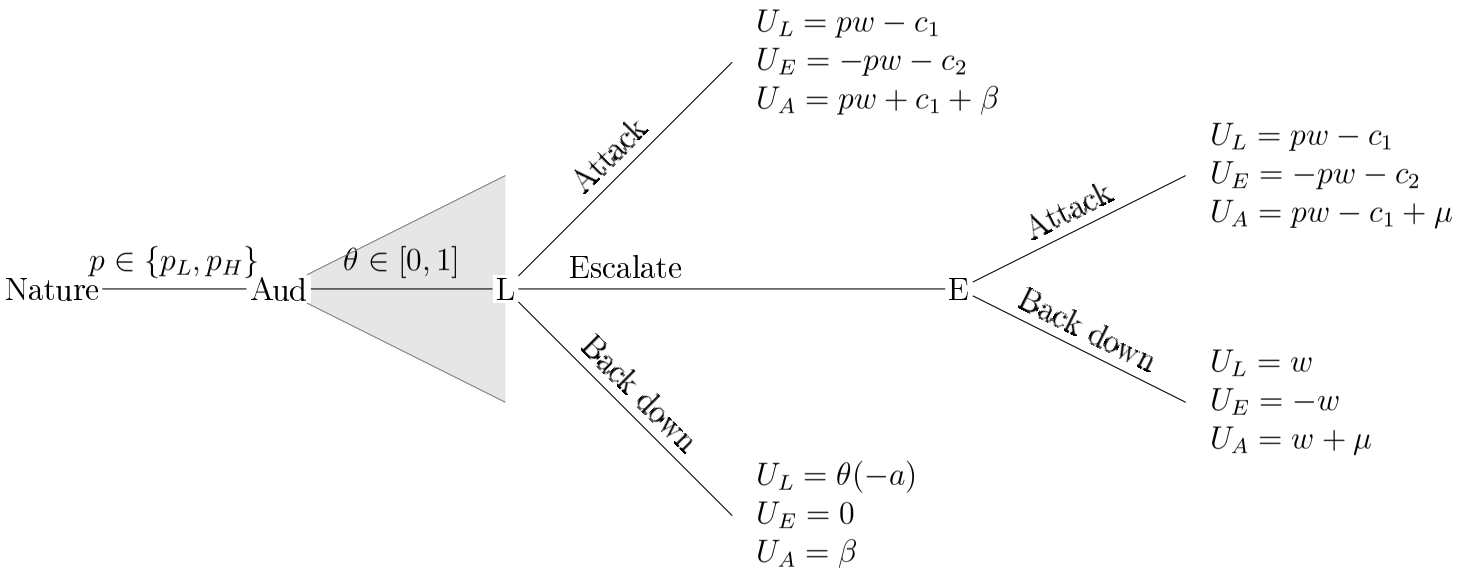

Figure 1: Stylized Model of Audience Costs With Reputational Concern

a

[a]Note that β represents the updated posterior belief about the home country's true type given the leader's actions.

References

[1] Beardsley, Kyle C. 2010. "Pain, Pressure and Political Cover: Explaining Mediation Incidence." *Journal of Peace Research* 47(4): 395-406.

[2] Downes, Alexander B. and Todd S. Sechser. 2010. "The Illusion of Democratic Credibility." *unpublished manuscript.*

[3] Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88(3): 577-592.

[4] Fearon, James D. 1997. "Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs." *The Journal of Conflict Resolution* 41(1): 68-90.

[5] Gelpi, Christopher and Michael Griesdorf. 2001. "Winners of Losers? Democracies in International Crisis, 1918-94." *American Political Science Review* 95(3): 633-647.

[6] Kurizaki, Shuhei. "Efficient Secrecy: Public versus Private Threats in Crisis Diplomacy." *American Political Science Review* 101(3) (2007): 543-58.

[7] Partell, Peter J. and Glenn Palmer. 1999. "Audience Costs and Interstate Crises: An Empirical Assessment of Fearon's Model of Dispute Outcomes." *International Studies Quarterly* 43(2): 389-405.

[8] Ramirez, Shawn L. "Diplomatic Options in the Shadow of An Audience: The Benefits of Private Mediation." *unpublished manuscript.*

[9] Ramsay, Kristopher W. "Politics at the Water's Edge." *Journal of Conflict Resolution* 48(4), 2004: 459-486.

[10] Schultz, Kenneth. 2001. *Democracy and Coercive Diplomacy* New York: Cambridge University Press.

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

[12] Slantchev, Branislav. "Audience Cost Theory and Its Audiences." *unpublished manuscript.*

[13] Slantchev, Branislav L. 2006. "Politicians, the Media, and Domestic Audience Costs." *International Studies Quarterly* 50: 445-477.

[14] Snyder, Jack and Erica Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *unpublished manuscript.*

[15] Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61(4): 821-840.

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