ISSN 2279-9362

Collegio Carlo Alberto

A Behavioral Foundation for Audience Costs

Avidit Acharya
Edoardo Grillo

No. 468
November 2016

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

A Behavioral Foundation for Audience Costs*

Avidit Acharya† Edoardo Grillo‡

August 2016

Abstract

We provide a behavioral foundation for audience costs by augmenting the standard crisis bargaining model with voters who evaluate material outcomes relative to an endogenous reference point. They vote to re-elect their leader when their payoff is high, and vote to replace him when their payoff is low. Politicians value re-election but also care about the outcome of the crisis. Backing down after a challenge may be costly to a leader because initiating the challenge has the potential to raise voters' expectations about their final payoff, generating the possibility that they suffer a payoff loss from disappointment when their leader backs down. Whether it is costly or beneficial to back down from a threat—and just how costly or beneficial it is—depends on the value of the reference point, which is determined endogenously in equilibrium.

Key words: crisis bargaining, audience costs, reference-dependent utility

*We thank Jim Fearon, Kris Ramsay and Ken Schultz for helpful conversations.

†Assistant Professor of Political Science, Stanford University, Encina Hall West, Room 406, Stanford CA 94305-6044 (email: avidit@stanford.edu).

‡Assistant Professor of Economics, Collegio Carlo Alberto, Via Real Collegio, 30, 10024 Moncalieri (Torino), Italy (email: edoardo.grillo@carloalberto.org).

1 Introduction

The possibility that citizens punish their politicians for backing down from an inter-state dispute has captured the interest of international relations scholars for more than two decades. In the original article on the topic, Fearon (1994) referred to the cost incurred by the politician as a result of this punishment as an "audience cost." By suggesting that the incentives that citizens provide to their politicians matter in understanding the choices leaders make in inter-state politics, his article generated substantial enthusiasm for a new line of research that connected international relations theory to domestic politics. Despite this, not many papers have explored the possible reasons for why citizens might punish their leaders this way, and how audience costs emerge from a theory that explicitly models voter preferences.[1]

In this paper, we consider one possible explanation for audience costs that is rooted in behavioral psychology. When rational politicians issue threats, voters may raise their expectations about how successful the politician will be in extracting concessions from the adversary. If the politician eventually backs down from the threat, voters may become disappointed. When voting behavior is based on this disappointment such that the politician's re-election probability is decreasing in the level of overall disappointment, then the politician suffers an audience cost from making the challenge and subsequently backing down. To explore this possibility, we augment the standard crisis bargaining model by adding a stage in which behavioral voters vote to re-elect or replace the incumbent politician.

The stylized crisis bargaining model that we extend is depicted in Figure 1. In the model the leaders of two countries, a potential Challenger, C, and a Defender, D, make sequential decisions. The leader of C is one of two types— weak, W, or strong, S— and type is private information, with the prior probability of the strong type denoted $q \in (0, 1)$. The leader of C first chooses whether or not to challenge D for a piece of territory that both value at $v > 0$. If C challenges, then the leader of D decides whether to resist or concede the territory. Finally, if D resists then C's leader can either escalate to war or back down.

If there is no challenge, then the territory remains with country D, which means that both types of C's leaders get a payoff of 0 while D gets a payoff of v. If the game ends with D's leader conceding, then the territory goes to country C, which results in both types of country C's leaders getting a payoff of v and D getting a payoff of 0. If the

[1] We review the few papers that have explored this possibility in Section 4.2.

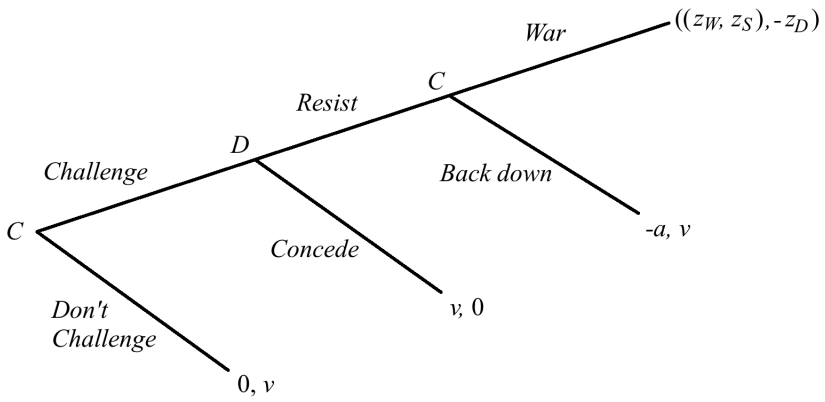

Figure 1

game ends with war, then the weak leader of C obtains a payoff z_W while the strong leader gets z_S and D gets $-z_D$. If the game ends with C's leader backing down, then the payoffs are $-a$ for C and v for D. The interpretation of this is that the territory remains with country D, and if $a > 0$ then C's leader incurs a cost from backing down in comparison to the payoff from not challenging to begin with. This is what Fearon (1994) calls the "audience cost." Although it is exogenous, he postulates an extension of the model in which it results from the possibility that the citizens of C punish their leader for backing down from the initial challenge.

To provide a behavioral foundation for this cost, we set the exogenous audience cost to $a = 0$ and augment the model above with voters who have reference-dependent payoffs in which the reference point is determined endogenously as in Kőszegi and Rabin (2006, 2007). The voters are citizens of C, and are of two types: hawks, whose material payoffs equal the crisis bargaining payoffs of the strong type of leader, and doves, whose material payoffs equal the crisis bargaining payoffs of the weak type of leader. If voters vote to re-elect their politician when their payoff is high, and vote to replace him when their payoff is low, then backing down after a challenge may be costly to the politician if the politician also values re-election. In particular, if voters are predominantly hawks, then entering a crisis by challenging the territory has the potential to raise the voters' reference points. Backing down could then generate a payoff loss due to disappointment. Citizens cannot be disappointed if the politician does not challenge the territory since this would not raise their expectations in the first place. Whether or not it is costly

to back down, however, and just how costly it is, depends on the value of the reference point, and whether voters are predominantly hawks or doves. If sufficiently many voters are sufficiently hawkish about war, then backing down after a challenge generates an audience cost. If sufficiently many voters are sufficiently dovish about war, then backing down generates an audience benefit. The reason is that after a challenge, the doves form the pessimistic expectation that they are likely to go to war, which they would like to avoid. If the politician backs down, these voters get a payoff gain from the sense of relief that war did not ensue. In the model, disappointment and relief are two psychological states that occur on the two opposite sides of an endogenous reference point.

After the pioneering work of Kahneman and Tversky (1979), numerous empirical and experimental studies have found evidence that individual behavior is consistent with the maximization of reference dependent payoffs.[2] The mounting evidence for reference dependent utility theory has also motivated several political science papers prior to ours. Levy (1997) reviews some of the early applications of prospect theory in international relations, and suggests areas of application. Other recent contributions in political science that are particularly relevant to our application here include Kimball and Patterson (1997) and Waterman et al. (1999) who show that the attitudes of voters towards an elected official are affected by their expectations, Alesina and Passarelli (2014) and Lockwood and Rockey (2016) who show that reference dependent preferences can explain departures from the predictions of standard voting models, Passarelli and Tabellini (2015) who show that reference dependent utility theory can explain political unrest, Grillo (2016) who uses reference dependent utility to provide a rationale for why voters believe the campaign promises of politicians, and Martin (2016) who shows that loss aversion explains why higher tax payers are more willing to punish corruption.

This paper has four more sections. Section 2 presents the model and Section 3 presents the equilibrium analysis. In Section 4, we discuss how some of the recent empirical evidence on audience costs squares with the predictions of our model. We also discuss other approaches that provide foundations for audience costs, and explain how our approach differs from these. Section 5 concludes.

[2]For example, in some relatively recent work, Fehr et al. (2011) find evidence for reference dependent payoffs in contractual relationships, Farber (2008) finds evidence for reference-dependent payoffs in labor markets, and Pope and Schweitzer (2011) find evidence for loss aversion in non-market settings.

2 Model

2.1 The Workhorse Model as a Benchmark

We start by introducing a set of standard assumptions on the workhorse crisis bargaining model described above (and depicted in Figure 1) and reporting its equilibria.

First, we assume that $v > z_S > 0 > z_W$ so that absent any audience cost or benefit, the strong type would choose war at his final decision node and the weak type would back down; and, the value of the territory for both types of C's leaders is greater than the payoff from war. Second, we assume that $-z_D < 0$ so that D would prefer to concede the territory than to go to war. These reduced form war payoffs can be interpreted as expected payoffs when war is costly and the outcome of war is uncertain.[3] Finally, to avoid having to deal with trivial sources of multiplicity arising from knife-edge cases of indifference, we maintain the assumption throughout the paper that the war payoffs z_W, z_S and $-z_D$ are all generic.

The assumption of generic war payoffs implies that the following three cases are exhaustive: (i) $-a > z_S > z_W$, (ii) $z_S > z_W > -a$ and (iii) $z_S > -a > z_W$.

In the first case, both the strong and weak leaders of country C back down at their final decision nodes, so D resists. Furthermore, since $z_S > 0$, this case can arise if and only if $a < 0$. As a result, there is a unique equilibrium in which both types of C challenge with certainty.

In the second case, the audience cost a is so high that both the weak and strong types of C choose war over backing down. Since $-z_D < 0$, there is a unique equilibrium in which D concedes and, consequently, both types of C challenge.

In the third case, the strong type of C chooses war at its final decision node, while the weak type backs down. If $q > v/(v + z_D)$, then there is a unique equilibrium in which D concedes, and both the strong and weak types of C challenge. On the other hand, suppose that $q \leq v/(v + z_D)$. Then, in equilibrium, D resists with probability $\min\{1, v/(a + v)\}$ and the strong type of C challenges. The weak type challenges with probability $qz_D/(1 - q)v$ if $a > 0$, with probability 1 if $a < 0$, and with any probability

[3]For example, the probability that country C wins the war is p. The cost of war incurred by the leader of country D is c_D, so the expected payoff for that leader is $-z_D := (1 - p)v - c_D$. The cost of war incurred by the weak leader of country C is c_W while the cost incurred by the strong leader is c_S. Thus, expected payoff under war for the weak leader of country C is $z_W := pv - c_W$ while the expected payoff for the strong leader of country C is $z_S := pv - c_S$. The assumptions that we make on z_W, z_S and z_D can then be translated to assumptions on p, c_W, c_S and c_D.

$\sigma_W \geq qz_D/(1-q)v$ if $a = 0$. In this latter case, there is a continuum of equilibria, including one in which the weak type challenges the territory with certainty.[4]

2.2 Augmenting the Model with Behavioral Voters

Our purpose here is to augment the workhorse model in such a way that any cost that the leader of C suffers from backing down arises *endogenously*.

To this end, suppose that there are no exogenous audience costs, $a = 0$, and that after the leaders of C and D make their decisions, a continuum of citizens of C cast votes to re-elect or replace their leader. The politician is re-elected if and only if a majority of voters vote to re-elect him. If the politician is not re-elected, he obtains only a payoff equal to the payoff that he gets from the crisis bargaining game. If he is re-elected, then he gets an additional payoff that we normalize to 1.

We consider the voters to be mechanical actors (whose behavior we specify below) so they are not players in the game.[5] Thus, an equilibrium of the game will specify only the behavior and beliefs of the leaders of the two countries. Let σ_θ denote the equilibrium probability with which the type $\theta \in \{W, S\}$ leader of country C challenges, σ_θ^w the equilibrium probability with which this type chooses war, and σ_D the equilibrium probability with which D resists. The equilibrium strategy profile is therefore $\sigma = \langle(\sigma_\theta, \sigma_\theta^w)_{\theta=W,S}, \sigma_D)\rangle$. Let $\tilde{q}$ denote the equilibrium posterior probability with which the leader of C is considered to be the strong type after choosing to challenge the territory. D's belief about C's type matters only at the information set at which D's chooses to resist or concede, so we may write an equilibrium to be simply the pair $\rho = (\sigma, \tilde{q})$.

There are two types of voters: those whose material payoffs are given by the payoffs of the strong type of leader of country C in the crisis bargaining game, and those whose material payoffs are given by the payoffs of the weak type in the same game. Fraction λ of voters are of the former type, while $1 - \lambda$ are of the latter type. We will refer to these two types of voters as hawks and doves respectively, and use the labels S and W for voters as well. Voters also have a psychological component of payoffs, which is reference-dependent. The material and psychological components of voter payoffs are

[4] In the cases where a type is indifferent between challenging and not challenging because $a = 0$, challenging with certainty is weakly dominant. So weak dominance as a criterion for equilibrium selection would select the equilibrium in which that type challenges.

[5] There are a continuum of them so any voting rule could have been selected in a model in which the voters are considered to be players. Further, we will assume below that voting is probabilistic so the assumption that voters are mechanical is standard.

additively separable for each type θ. We write the sum of these two components as

$$u_\theta = \pi_\theta + \eta(\pi_\theta - \mathbb{E}^\rho[\pi_\theta|\mathcal{I}]), \qquad \theta \in \{W, S\} \tag{1}$$

where π_θ is the material payoff of the type θ leader in the crisis bargaining game, $\mathbb{E}^\rho[\cdot|\mathcal{I}]$ denotes the expectation operator evaluated at an information set $\mathcal{I}$, and given an equilibrium of the game ρ, and $\eta \geq 0$ is the weight on the psychological component of payoffs. We assume that the voters' reference points are determined at the information sets that arise immediately after the initial choice of C's leader to challenge or not challenge, which we label $\mathcal{I}_{ch}$ and $\mathcal{I}_d$ respectively. This assumption reflects the salience of C's initial decision in forming voter expectations. At the information set $\mathcal{I}_d$ the voter knows that her payoff will be 0, so $\mathbb{E}^\rho[\pi_\theta|\mathcal{I}_d] = 0$ for both $\theta \in \{W, S\}$ and all equilibria ρ. Finally, we assume that voters share D's belief about C's type at $\mathcal{I}_{ch}$: the posterior probability with which they think that C is strong is also $\tilde{q}$.

Voting is probabilistic. Each voter receives a stochastic preference shock ε that is drawn uniformly from the interval $\left[-\frac{1}{2\alpha}, \frac{1}{2\alpha}\right]$ and independently across voters, and each voter votes to reelect the incumbent politician if and only if his payoff (the deterministic part plus the shock) exceeds a stochastic threshold u that is drawn uniformly from the interval $\left[-\frac{1}{2\beta}, \frac{1}{2\beta}\right]$. Here, β measures the overall responsiveness of the electorate to the outcome of the crisis. We make the standard assumption that α and β are sufficiently small so that the probability that the politician is re-elected is

$$\frac{1}{2} + \beta\big[\lambda u_S + (1-\lambda)u_W\big] \tag{2}$$

This quantity is also the additional expected payoff that the leader gets due to the fact that he may be re-elected. Since the term in squared brackets of this expression is simply the population-weighted (utilitarian) average of voters' payoffs, the leader of country C maximizes a payoff equal to the payoff that he receives in the crisis bargaining game, which depends on his type, plus β times the utilitarian average of voters' payoffs, which includes both the material and psychological parts.

Remark 1 There are other assumptions that could give rise to the result that C's leader maximizes the payoff from the crisis plus (2). One is to directly assume that the leader of C has weighted utilitarian preferences and places weight β on the average voter payoff. Another is to say that C's citizens have non-electoral ways of rewarding and

punishing their leader such that the leader internalizes the average voter payoff. Under such assumptions, our results are also applicable to cases, such as dictatorships, where leaders are not directly chosen by voters.[6] However, under these alternative assumptions, β would no longer be a measure of the responsiveness of the electorate to the outcome of the crisis. Instead, it would measure the extent to which the politician's payoffs were other-regarding, or the extent to which the political process generated incentives for politicians to consider the average voter's interests.

Remark 2 For the assumption of endogenous reference-dependent payoffs to play a role in affecting equilibrium behavior, it must be that the endogenous reference point is not updated at every information set.[7] The assumption that voters update their reference point at the two information sets that arise after C's initial choice is natural in our application, capturing the idea that citizens form their expectations based on what they learn after observing their own leader's initial policy choice, but not on the details of inter-state crisis bargaining, which, in practice, are typically opaque. That said, there does not yet exist a theory about how to select the information sets at which the endogenous reference points are updated in sequential move games. Given this, in the Supplemental Appendix we discuss the equilibrium consequences of choosing other sets of information sets in which the reference point is updated.

2.3 Endogenous Payoffs and Equilibrium Definition

Substituting (1) into (2) and simplifying, the politician's probability of re-election is

$$\frac{1}{2} + \beta \left[\pi^{\lambda} + \eta \left(\pi^{\lambda} - \mathcal{R}^{\rho}[\mathcal{I}] \right) \right] \tag{3}$$

where

$$\pi^{\lambda} := \lambda \pi_S + (1 - \lambda) \pi_W \tag{4}$$

is the population-weighted average of material payoffs given any outcome of the crisis bargaining game, and

$$\mathcal{R}^{\rho}[\mathcal{I}] := \lambda \mathbb{E}^{\rho}[\pi_S | \mathcal{I}] + (1 - \lambda) \mathbb{E}^{\rho}[\pi_W | \mathcal{I}] \tag{5}$$

[6] Weeks (2008) argues that audience costs can arise even in dictatorships.

[7] If voters updated their reference point at all information sets, then the equilibrium of our model would be behaviorally identical to the perfect Bayesian equilibrium of the same game without reference-dependent payoffs, i.e. with $\eta = 0$. See the Supplemental Appendix.

is the the population weighted average value of the endogenous reference point evaluated at an equilibrium ρ and information set $\mathcal{I}$.

As mentioned above, for both types of voters, $\theta \in \{W, S\}$, the endogenous reference point in the case where the C's leader chooses to not challenge the territory is $\mathbb{E}^{\rho}[\pi_{\theta}|\mathcal{I}_d] = 0$. This implies that $\mathcal{R}^{\rho}[\mathcal{I}_d] = 0$ independently of the equilibrium ρ. Therefore, if either type of leader chooses not to challenge the territory, he is re-elected with probability $\frac{1}{2}$ and obtains a payoff of $\frac{1}{2}$ from not challenging.

At the information set $\mathcal{I}_{ch}$, voters observe C's leader challenged. Thus, the endogenous reference point of a type θ voter after a challenge is

$$\mathbb{E}^{\rho}[\pi_{\theta}|\mathcal{I}_{ch}] = (1-\sigma_D)\, v + \sigma_D \big[\tilde{q}(\sigma_S^w z_{\theta} + (1-\sigma_S^w)0) + (1-\tilde{q})(\sigma_W^w z_{\theta} + (1-\sigma_W^w)0)\big] \quad (6)$$

The reference point is, therefore, a weighted average of the payoffs that arise at the terminal nodes following the initial challenge, with weights given by the probabilities with which these nodes are reached given the voters' equilibrium beliefs.

This implies that the population weighted average value of the endogenous reference point after country C's leader challenges the territory is

$$\mathcal{R}^{\rho}(\mathcal{I}_{ch}) = (1-\sigma_D)\, v + \sigma_D \left[\tilde{q}\sigma_S^w + (1-\tilde{q})\sigma_W^w\right] z^{\lambda} \quad (7)$$

where $z^{\lambda} = \lambda z_S + (1-\lambda) z_W$, which follows from (4). Therefore, if the game ends with country D conceding, the payoff to both types of country C's leaders is

$$v + \frac{1}{2} + \beta\Big[v + \eta\big(v - \mathcal{R}^{\rho}(\mathcal{I}_{ch})\big)\Big] \quad (8)$$

If the game ends with C's leader backing down, the expected payoff to both types of C's leaders is

$$0 + \frac{1}{2} + \beta\Big[0 + \eta\big(0 - \mathcal{R}^{\rho}(\mathcal{I}_{ch})\big)\Big] \quad (9)$$

and if the game ends with war, the expected payoff to each type θ of C's leaders from choosing war is

$$z_{\theta} + \frac{1}{2} + \beta\Big[z^{\lambda} + \eta\big(z^{\lambda} - \mathcal{R}^{\rho}(\mathcal{I}_{ch})\big)\Big] \quad (10)$$

The payoffs from the various outcomes of the game to the leader of D are simply D's payoffs in the crisis bargaining game, depicted in Figure 1.

Since the payoffs to the two types of leaders of country C are endogenous to the equilibrium strategy and beliefs of D, we say that $\rho = (\sigma, \tilde{q})$ is an equilibrium of the

model if (i) $\tilde{q}$ is consistent with Bayesian updating given σ, and (ii) no type of either player has a profitable deviation from the strategy profile σ given beliefs $\tilde{q}$ when the payoffs to all of the outcomes of the game are computed at the equilibrium ρ. In this sense, an equilibrium of a model in which players have reference dependent preferences with endogenous reference points has the fixed point characteristic that is typical of a rational expectations equilibrium: the reference points are derived from equilibrium behavior and equilibrium behavior is consistent with the endogenous references points.

2.4 Endogenous Audience Costs

It is now already apparent that the politician may suffer an endogenous audience cost from backing down after making a threat. The cost for the leader of C from backing down after a challenge is the payoff difference from backing down after a challenge and not challenging at the start of the game, which is

$$a_s^\rho = \beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch}) \tag{11}$$

If this quantity is positive, it represents an endogenous audience cost that is sunk the moment C's leader decides to challenge. For this reason, we refer to a_s^ρ as a *sunk audience cost* if it is positive, or benefit if it is negative.

Similarly, the payoff difference between going to war and backing down for a leader of type θ is $z_\theta + \beta(1+\eta)z^\lambda$. This exceeds the same payoff difference in the workhorse model without voters by the quantity

$$a_t = \beta(1+\eta)z^\lambda \tag{12}$$

Therefore, if there are sufficiently many hawks in the population so that z^λ is positive then the leader of C has an extra incentive to go to war over backing down. In particular, electoral incentives can commit even the weak type of politician to war.[8] On the other hand, if there are sufficiently many doves in the population then z^λ will be negative, so electoral incentives can commit even the strong type politician to back down rather than choose war. For this reason, we refer to a_t as a *tying-hands audience cost* or benefit.

[8]Even though $z_W < 0$, it is possible that $z_W + \beta(1+\eta)z^\lambda > 0$ so that the weak leader's payoff in (10) is greater than his payoff in (9). This means that in the augmented model with electoral incentives, even a weak type may choose war over backing down. See the analysis of case (ii) in Section 3.

Our terminology is consistent with Fearon's discussion of two kinds (or functions) of audience costs: sunk costs and tying hands costs (Fearon, 1997).

Both the sunk and tying hands audience costs are endogenous quantities, but the sunk audience cost is also an equilibrium quantity since it depends on $\mathcal{R}^{\rho}(\mathcal{I}_{ch})$, which is an equilibrium quantity. In fact, the sign of a_s^{ρ} is completely determined by the sign of $\mathcal{R}^{\rho}(\mathcal{I}_{ch})$, so whether there is an audience cost or benefit is also determined in equilibrium. In addition, the assumption that voters have reference dependent payoffs is necessary for generating a sunk audience cost since $a_s^{\rho} = 0$ if $\eta = 0$. But this assumption is not necessary for generating a tying hands cost since $\eta = 0$ does not imply that $a_t = 0$.[9]

3 Equilibrium

As in the benchmark model, we have three cases to consider: (i) $-a_t > z_S > z_W$, (ii) $z_S > z_W > -a_t$, and (iii) $z_S > -a_t > z_W$.

In the first case, both the strong and weak types back down, so there is a double continuum of equilibria in which D resists and each type of C challenges with any probability. Since $z_S > 0 > z_W$, this case arises only when there are sufficiently many doves in the population, so that $z^{\lambda} < 0$.

In the second case, both types of leaders choose war at their final decision nodes. Therefore, in the unique equilibrium, D concedes and both types challenge. This case arises only if there are sufficiently many hawks in the population, so that z^{λ} is sufficiently high. Unlike in the first case, in this case the electorate works as a commitment device that enables C to credibly threaten to escalate the crisis to war, forcing D to concede.

We analyze the third case below.

3.1 Analysis

Since we pinned down equilibrium behavior for cases (i) and (ii) above, in this section we analyze case (iii), where $z_S > -a_t > z_W$. In this case, the weak and strong types make separating choices at their final decision nodes: the weak type chooses to back

[9]Since the tying hands audience cost is not an equilibrium quantity, our model says that a leader cannot strategically generate commitment to one of the two actions at the final decision nodes of the game. Instead, the magnitude and sign of the tying-hands audience cost is determined directly by the model's fundamental parameters (the responsiveness of voters to the outcome of the crisis, β; the weight on the psychological part of their payoffs, η; and the population average of material war payoffs, z^{λ}).

down while the strong type chooses war. As a result, we have

$$\mathcal{R}^{\rho}(\mathcal{I}_{ch}) = (1-\sigma_D)v + \sigma_D \tilde{q} z^{\lambda}. \tag{13}$$

From here, our analysis proceeds in two steps. In the first step, we prove that in any equilibrium the strong type of C challenges with probability 1. In the second step, we provide a characterization of the full equilibrium set by searching for equilibria in three exhaustive cases: the case where D concedes, the case where D resists, and the case where D mixes between conceding and resisting.

In equilibrium, the strong type challenges Suppose, for the sake of contradiction, that there is an equilibrium in which the strong type of C challenges with probability less than 1. If there were such an equilibrium, then the strong type's expected payoff from challenging could not exceed his expected payoff from not challenging, i.e.

$$\begin{aligned} 0 &\geq \sigma_D\big[z_S + \beta\big(z^{\lambda} + \eta(z^{\lambda} - \mathcal{R}^{\rho}(\mathcal{I}_{ch}))\big)\big] + (1-\sigma_D)\big[v + \beta\big(v + \eta(v - \mathcal{R}^{\rho}(\mathcal{I}_{ch}))\big)\big] \\ &= \sigma_D\big[z_S + \beta(1+\eta(1-\tilde{q}))z^{\lambda}\big] + (1-\sigma_D)(1+\beta)v \end{aligned} \tag{14}$$

where the second line follows from substituting $\mathcal{R}^{\rho}(\mathcal{I}_{ch})$ from (13). If $z^{\lambda} \geq 0$, the right side of (14) is strictly positive, establishing the contradiction. If $z^{\lambda} < 0$, we have $z_S + \beta(1+\eta(1-\tilde{q}))z^{\lambda} > z_S + \beta(1+\eta)z^{\lambda} > 0$, where the last inequality follows from the fact that we are analyzing a case where $z_S > -a_t$, and by the definition of a_t. Thus, (14) is again positive, establishing the contradiction. □

Since the strong type always challenges in equilibrium, $\tilde{q}$ is pinned down by Bayes rule. In particular,

$$\tilde{q} = \frac{q}{q + \sigma_W(1-q)}. \tag{15}$$

Then, given that the two types of C separate at their final decision nodes, D chooses to concede only if

$$\tilde{q}(-z_D) + (1-\tilde{q})v \tag{16}$$

is weakly less than 0, in which case D's expected payoff from resisting is weakly lower than her expected payoff from backing down. D chooses to resist only if (16) is weakly greater than 0, and D chooses to mix between conceding and resisting only if it is exactly equal to 0. We now complete the characterization of the equilibrium set, organizing the analysis according to D's equilibrium choices.

Equilibria where D concedes Suppose that D concedes, so $\sigma_D = 0$. Then $\mathcal{R}^\rho(\mathcal{I}_{ch}) = v$ and the payoff to both types of C from challenging is $v + \frac{1}{2} + \beta v$, which is greater than $\frac{1}{2}$, the payoff from not challenging. So both types challenge, and $\tilde{q} = q$. Then, for it to be optimal for D to concede we need (16) to be at least as large as 0 when $\tilde{q} = q$; that is, we need $q \geq v/(v + z_D)$. Thus, if the prior q is above $v/(v + z_D)$ there is an equilibrium in which D concedes, and both types of C challenge. If $q < v/(v + z_D)$, then D has a profitable deviation and there is no equilibrium in which D concedes for sure. □

Equilibria where D resists Next, consider the case where D resists, so $\sigma_D = 1$. For D to want to resist we would need (16) to be weakly greater than 0 evaluated when $\tilde{q}$ is given by (15). Thus, we need $\sigma_W \geq qz_D/(1-q)v$. This latter inequality defines a feasible value of σ_W if and only if $q \leq v/(v + z_D)$.

Now suppose that $\eta = 0$. The weak type of C is always indifferent between challenging and not challenging since his expected payoff from challenging is $\frac{1}{2} - \beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch}) = \frac{1}{2}$ and his expected payoff from not challenging is also $\frac{1}{2}$. Therefore, when $q \leq v/(v + z_D)$ and $\eta = 0$, there is a continuum of equilibria in which D resists and the weak type of C challenges with any probability $\sigma_W \geq qz_D/(1-q)v$.

Lastly, consider the case where $\eta > 0$. If $z^\lambda > 0$, then there is no equilibrium where D resists, because if this were the case, the weak type's payoff from not challenging, $\frac{1}{2}$, would exceed his equilibrium payoff from challenging, $\frac{1}{2} - \beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch})$, giving this type a profitable deviation. On the other hand, if $z^\lambda < 0$ then the weak type would want to challenge. Therefore, when $\eta > 0$ there is an equilibrium in which D resists if and only if $z^\lambda < 0$. In this equilibrium, both the weak and strong types challenge. □

Equilibria where D mixes Suppose that D mixes between conceding and resisting. To mix, D must be indifferent, so (16) must equal 0. Substituting (15) into this indifference condition gives us

$$0 = \frac{q}{q + \sigma_W(1-q)}(-z_D) + \frac{\sigma_W(1-q)}{q + \sigma_W(1-q)}v \tag{17}$$

This pins down the equilibrium value of σ_W, which is $\sigma_W = qz_D/(1-q)v$. As in the previous case, this condition defines a feasible value for σ_W if and only if $q \leq v/(v + z_D)$. If this condition is satisfied, then $\tilde{q} = v/(v + z_D)$. Otherwise, there is no equilibrium in

which D mixes. Thus, suppose $q < v/(v + z_D)$.[10] Since $\sigma_W \in (0, 1)$, the weak type of C must also be indifferent between challenging and not challenging, we need

$$0 = (1 - \sigma_D)\left(v + \beta\left[v + \eta\left(v - \mathcal{R}^\rho(\mathcal{I}_{ch})\right]\right) + \sigma_D\left(-\beta\eta\mathcal{R}^\rho(\mathcal{I}_{ch})\right) \tag{18}$$

Now we substitute the equilibrium belief $\tilde{q} = v/(v + z_D)$ into $\mathcal{R}^\rho(\mathcal{I}_{ch})$ in (13), and then $\mathcal{R}^\rho(\mathcal{I}_{ch})$ from (13) into (18), and solve for σ_D to get

$$\sigma_D = \frac{(1+\beta)v}{(1+\beta)v + \beta\eta\tilde{q}z^\lambda} = \frac{(1+\beta)(v+z_D)}{(1+\beta)(v+z_D) + \beta\eta z^\lambda} \tag{19}$$

This implies that there is no equilibrium in which D mixes if $z^\lambda < 0$ or $\eta = 0$, but there is such an equilibrium when z^λ and η are both positive. □

3.2 Summary

We summarize the main findings of our analysis in Proposition 1 below. Since the sunk audience cost a_s^ρ is an equilibrium quantity, we also report its equilibrium value. The proof of the proposition is in the text above.

Proposition 1

(i) If $-a_t > z_S > z_W$ then there is a double continuum of equilibria in which both the strong and weak types of C back down, D resists, and both types of C are indifferent between not challenging and challenging, so each may challenge with any probability. In all of these equilibria, $a_s^\rho = 0$.

(ii) If $z_S > z_W > -a_t$ then there is a unique equilibrium in which both types of C choose war, D concedes, and both types of C challenge, so $a_s^\rho = \beta\eta v$.

(iii) If $z_S > -a_t > z_W$ then in any equilibrium, the strong type of C chooses war and challenges, while the weak type backs down. In addition, if $q > v/(v + z_D)$, then there is a unique equilibrium in which D concedes and the weak type of C also challenges, so again $a_s^\rho = \beta\eta v$. If $q < v/(v + z_D)$ then we have three subcases:

[10] The case where $q = v/(v + z_D)$ would yield a continuum of equilibria. Since our assumption that z_D is generic rules out this case, we do not characterize the set of equilibria for this case.

(a) If $\eta = 0$, then there is a continuum of equilibria in which D resists, and the weak type of C challenges with any probability $\sigma_W \geq qz_D/(1-q)v$. In all of these equilibria, $a_s^\rho = 0$, so there is no sunk audience cost or benefit.

(b) If $\eta > 0$ and $z^\lambda < 0$, there is a unique equilibrium in which D resists and the weak type of C challenges, so there is a sunk audience benefit, $a_s^\rho = \beta\eta q z^\lambda < 0$.

(c) If $\eta > 0$ and $z^\lambda > 0$ there is a unique equilibrium in which D resists with probability

$$\sigma_D = \frac{(1+\beta)(v+z_D)}{(1+\beta)(v+z_D) + \beta\eta z^\lambda}$$

and the weak type of C challenges with probability $\sigma_W = qz_D/(1-q)v$. In this case, the sunk audience cost is

$$a_s^\rho = (1-\sigma_D)\left[1+\beta(1+\eta)\right]v = \frac{\beta\eta z^\lambda}{(1+\beta)(v+z_D) + \beta\eta z^\lambda}\left[1+\beta(1+\eta)\right]v$$

Therefore, the equilibrium payoffs of the augmented model are unique and the equilibria are analogous to the equilibria of the benchmark model.

In the augmented model, the tying hands audience cost, a_t, defines the threshold that separates the three cases where the weak and strong types both back down, both choose war, and make different choices at their final decision nodes, exactly as the exogenous audience cost a does in the benchmark model.

By comparing the equilibrium predictions of the augmented model with those of the benchmark model, we observe that in case (i) the micro-foundation of audience costs enlarges the set of equilibria by allowing C's leaders to challenge with probability lower than 1, while in case (ii) it delivers exactly the same predictions.

In case (iii), there is no sunk audience cost or benefit when $\eta = 0$. This establishes the necessity of reference dependent payoffs to produce a sunk audience cost in our setting. In this case, for each equilibrium of the augmented model, there is a behaviorally identical equilibrium of the $a = 0$ case of the benchmark model; and vice versa. When $\eta > 0$, the sign of z^λ determines whether there is a sunk audience cost or benefit. When there is a sunk audience benefit, equilibrium behavior in the augmented model is identical to equilibrium behavior in the benchmark model for the case of $a < 0$. When there is a sunk audience cost, equilibrium choices in the augmented model are also similar to equilibrium choices in the benchmark case for $a > 0$. The weak type of C mixes with the same probability in both models, but the probability with which D mixes is different, even

after accounting for the equilibrium value of the sunk audience cost. This is because the indifference condition that pins down D's mixing probability in the augmented model, equation (18), is qualitatively different from the analogous indifference condition in the benchmark model. One key difference is that the augmented model includes re-election payoffs that differ across terminal nodes. Another key difference is that the psychological part of voter payoffs that enters in C's payoff when D concedes also contains the sunk audience cost as a component, whereas in the benchmark model the exogenous audience cost a does not enter C's payoff when D concedes.

3.3 Comparative Statics

Since the tying hands and sunk audience costs are endogenous quantities in the augmented model, we can study their comparative statics.

We start by reporting the comparative statics of the tying hands audience cost, a_t. The sign of a_t is determined by the sign of z^λ, so there is an audience cost when citizens are predominantly hawks, and an audience benefit when they are predominantly doves. In addition, the magnitude of this cost or benefit is increasing in how predominant the hawks or doves are. The magnitude is also increasing in β, which means that when voting behavior is more responsive to the outcome of the crisis, or when the politician weights the average voter payoff more, there is a larger audience cost or benefit. Lastly, the magnitude is increasing in η, which means that when the psychological part of voter payoffs becomes more important, there is a greater audience cost or benefit. Therefore, the tying hands audience cost works as a commitment device that commits the politician to war when sufficiently many citizens are hawkish about war, when the outcome of the crisis matters more in their voting decisions, and when expectations matter more in determining their payoffs.

The sunk audience cost a_s^ρ is an equilibrium quantity that potentially varies with the equilibrium updated belief $\tilde{q}$ that C's leader is the strong type, the equilibrium probability σ_D with which D resists, and the equilibrium choices of C's types at their final decision nodes. So we must take this into account when studying the comparative statics of a_s^ρ. These comparative statics are by and large similar to those of the tying hands audience cost, with only a few notable differences. Again, the sign of a_s^ρ is determined by the sign of z^λ. As well, the magnitude of this audience cost is again increasing with the magnitude of z^λ. When $z^\lambda > 0$, the sunk audience cost a_s^ρ is increasing in both β and η. However, when $z^\lambda < 0$, it is piecewise constant in these parameters, with a

jump to 0 when $-a_t$ crosses z_S. This jump is positive if $q > v/(v + z_D)$ and negative if $q < v/(v + z_D)$. This means that when the citizens are predominantly doves, the audience benefit is weakly decreasing when the prior probability that C is the strong type is high, and weakly increasing when the prior is low. Finally, the magnitude of a_s^ρ is increasing in v. If voters are predominantly hawks, then as the value of the disputed territory goes up, the expected payoff of the voters after a concession by country D goes up as well. As a result, the reference point of the voters after a challenge, and hence the sunk audience cost, increases.[11]

4 Discussion

4.1 Empirical Evidence

Although the predictions of our model are mostly novel, some of our assumptions and predictions find support in experimental investigations of the audience cost.

The first experimental study of the audience cost was done by Tomz (2007), who estimates the sunk audience cost from survey data.[12] Tomz (2007) estimates a positive audience cost, and finds that the audience cost is higher among more politically active respondents. Though political engagement may not be the obvious way to measure voter responsiveness, this finding provides some evidence that is consistent with our prediction that the audience cost is increasing in the responsiveness of the electorate, β, to the outcome of the crisis.[13] That said, Tomz (2007) also presents some evidence that the public punishes leaders for bluffing because they think that bluffing hurts the leader's (and country's) reputation. While his evidence on the reputational harm of bluffing is different from the disappointment-based mechanism in our paper, this evidence is based on agents self-reporting their disapproval of bluffing, and these agents may not have had a consistent way of expressing their disappointment for the leader not following through on a challenge.

[11] This comparative static result with respect to v would continue to hold even if we substituted the payoffs following a war with the lottery payoffs described in footnote 3.

[12] Tomz (2007) argues that despite concerns about external validity, the experimental approach sidesteps several of the challenges in estimating the audience cost in observational studies, such as partial observability and strategic selection (Schultz, 2001).

[13] This result was replicated in the UK by Davies and Johns (2013), who found that the audience cost was highest among the most politically engaged British respondents. However, one result of theirs that goes against the grain of our predictions concerning the relationship between responsiveness and the audience cost is that political knowledge, which may also be correlated with responsiveness, did not substantially moderate the audience cost.

Building on Tomz's (2007) approach, Trager and Vavreck (2011) estimate a leader's public approval at every outcome of the crisis bargaining game, enabling them to estimate both the sunk audience cost as well as the tying hands audience cost as defined in this paper. They estimate positive values for both of these costs. They also find that presidential approval is highest when the adversary concedes, and can be lowest when the leader backs down —even lower than approval after the war outcome. They also show that among respondents who oppose a hawkish foreign policy rhetoric, there is an audience benefit rather than cost. Taking their presidential approval measure to be a proxy for the election payoff given in (2), these findings support our model. In particular, they support the prediction that the sign of z^λ determines the sign of the audience cost. If there are sufficiently many doves in the electorate, then there can be an audience benefit, though in their data Trager and Vavreck (2011) find that the hawks outnumber the doves.

Also building on Tomz (2007), Davies and Johns (2013) estimate the audience cost in the UK with variation in crisis type. They find that among voters, the disapproval for bluffing by the British prime minister was lower in a nuclear crisis scenario than in an ally defense crisis scenario, which was in turn lower than in a hostage crisis scenario. This suggests that the audience cost may potentially vary with the importance or scale of the issue, measured in our model by v. However, whether their findings show that it increases or decreases with scale remains unclear.

These studies provide some suggestive evidence for our theory, but more can be done to directly test the assumptions and predictions of our model.

4.2 Other Approaches

One influential theory of the audience cost is that leaders suffer a cost from the damage to their reputation that bluffing causes.[14] A simple and natural extension to the benchmark model that captures this story says that voters prefer to re-elect the strong type and replace the weak type. Suppose that the strong type challenges, and chooses war over backing down. If the weak type separates from the strong type at his final decision node, then he is not re-elected. However, he would not be re-elected even if he separated at the initial decision node, as this decision would also reveal his type to the voters. Therefore, this simple reputation-based extension does not produce an endogenous audience cost.

[14]As mentioned above, this is the theory that Tomz (2007) claims to find the strongest empirical support for based on open-ended survey responses.

Smith (1998) circumvents this problem by assuming that there are a continuum of types in an ally defense scenario. When the politician is inferred to be stronger, he is re-elected with higher probability. The set of types is partitioned into those that announce that they will support the ally against the adversary, and those that announce that they will not. In equilibrium, those that announce that they will support the ally follow through. If a deviation takes place, however, then Smith (1998) has the voters think that the type is the weakest possible type and re-elect him with the lowest probability. Thus, he generates audience costs with the help of off-path beliefs. However, for every profile of parameters (i.e. payoffs and initial beliefs) his game also has equilibria in which audience costs do not arise. Moreover, these equilibria cannot be ruled out using standard refinements.[15]

Guisinger and Smith (2002) also develop a theory of audience costs based on reputation, but depart further from the standard crisis bargaining scenario. In their model, two countries play a repeated demand bargaining game with adverse selection. In the one shot game, communicating a credible threat is not possible; but since the game is repeated, credible communication can be supported by an equilibrium strategy profile that reverts to babbling if the lying side is caught. Since payoffs are lower in the babbling equilibrium, voters would like to replace the lying politician after he is caught and start afresh with a new leader. Again, audience costs are supported by the selection of one of many possible equilibria of the game; and, in fact, equilibria that are renegotiation-proof in the sense of Farrell and Maskin (1989) do not support audience costs.

Other papers that provide foundations for audience costs include Ashworth and Ramsay (2010) and Slantchev (2006). Slantchev (2006) studies a game between a voter, politician, opposition party, and media, abstracting away from the foreign adversary. He shows that an audience cost for bad policies arises when the media can verify the opposition's claim that the policy is not going well. It is better to not pursue a policy that will fail because the media will then have no harmful evidence to report to voters. Ashworth and Ramsay (2010) take a mechanism design approach and show that an optimizing voter would design incentives to punish a politician for bluffing. However, their

[15]Smith's (1998) game is neither a standard signaling game, nor a standard cheap talk game, though it has some features of both. This means that the standard equilibrium refinements for signaling games must be adapted to his specific game. Furthermore, in his model, if the weakest possible type is better off by threatening the intervention and then not following through despite the bluff being called, types above it may also want to do the same. As a result, refinements like the ones proposed by Banks and Sobel (1987) do not uniquely select equilibria that support audience costs.

voter is only boundedly rational because he cannot condition the re-election probabilities on all aspects of the crisis bargaining outcome.

Our paper differs from the prior literature in at least three ways. First, we directly extend the canonical crisis bargaining model, endogenizing the audience costs in such a way as that equilibrium behavior in the extended model is directly analogous to equilibrium behavior in the benchmark model with exogenous audience costs. Second, we do this with behavioral voters. So, although we provide a micro-foundation for audience costs, our goal is not to rationalize these costs.[16] Third, and most importantly, our model provides a psychological theory for audience costs, based on disappointment and relief. The mechanism is new, and its implications can be tested empirically.

5 Conclusion

We have developed a new theory of audience costs based on disappointment and relief, by adding to the standard crisis bargaining model a voting stage in which voters have retrospective reference dependent payoffs. Our model endogenizes both the sunk audience cost and the tying hands audience cost, and produces new comparative statics predictions about the sign and magnitudes of these costs. If voters are predominantly hawkish about war, then both audience costs are positive but if they are predominantly dovish then both audience costs are negative, turning them into audience benefits. The magnitudes of these audience costs or benefits are increasing in the responsiveness of the electorate to the outcome of the crisis, as well as in the salience of the psychological component of payoffs. The magnitude of the sunk audience cost is also increasing in the value of the territory, or the importance of the issue to voters. These comparative statics predictions can be tested empirically.

[16] As Gul and Pesendorfer (2006) have shown, agents with endogenous reference dependent utility have preferences that violate transitivity.

References

Alesina, A. and F. Passarelli (2014): "Loss Aversion in Politics," *NBER Working Paper # 21077*.

Ashworth, S. and K. W. Ramsay (2010): "Should Audiences Cost? Optimal Domestic Constraints in International Crises," *manuscript, Princeton University*.

Banks, J. S. and J. Sobel (1987): "Equilibrium selection in signaling games," *Econometrica*, 647–661.

Davies, G. A. and R. Johns (2013): "Audience costs among the British public: the impact of escalation, crisis type, and prime ministerial rhetoric," *International Studies Quarterly*, 57, 725–737.

Farber, H. S. (2008): "Reference-dependent preferences and labor supply: The case of New York City taxi drivers," *The American Economic Review*, 98, 1069–1082.

Farrell, J. and E. Maskin (1989): "Renegotiation in repeated games," *Games and Economic Behavior*, 1, 327–360.

Fearon, J. D. (1994): "Domestic political audiences and the escalation of international disputes." *American Political Science Review*, 88, 577–592.

——— (1997): "Signaling foreign policy interests tying hands versus sinking costs," *Journal of Conflict Resolution*, 41, 68–90.

Fehr, E., O. Hart, and C. Zehnder (2011): "Contracts as Reference Points–Experimental Evidence," *American Economic Review*, 101, 493–525.

Grillo, E. (2016): "The hidden cost of raising voters expectations: Reference dependence and politicians credibility," *Journal of Economic Behavior & Organization*, 130, 126 – 143.

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

Gul, F. and W. Pesendorfer (2006): "The revealed preference implications of reference dependent preferences," *manuscript, Princeton University*.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–91.

Kőszegi, B. and M. Rabin (2006): "A model of reference-dependent preferences," *The Quarterly Journal of Economics*, 1133–1165.

——— (2007): "Reference-Dependent Risk Attitudes," *American Economic Review*, 97, 1047–1073.

Kimball, D. C. and S. C. Patterson (1997): "Living up to Expectations: Public Attitudes toward Congress," *Journal of Politics*, 59, 701–728.

Levy, J. S. (1997): "Prospect theory, rational choice, and international relations," *International Studies Quarterly*, 41, 87–112.

Lockwood, B. and J. Rockey (2016): "Negative Voters? Electoral Competition with Loss-Aversion," *manuscript, University of Warwick*.

Martin, L. (2016): "Taxation, loss aversion, and accountability: theory and experimental evidence for taxation?s effect on citizen behavior," *manuscript, Yale University*.

Passarelli, F. and G. Tabellini (2015): "Emotions and Political Unrest," *IGIER Working Paper # 474*.

Pope, D. G. and M. E. Schweitzer (2011): "Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes," *The American Economic Review*, 101, 129–157.

Schultz, K. A. (2001): "Looking for audience costs," *Journal of Conflict Resolution*, 45, 32–60.

Slantchev, B. L. (2006): "Politicians, the media, and domestic audience costs," *International Studies Quarterly*, 50, 445–477.

Smith, A. (1998): "International crises and domestic politics," *American Political Science Review*, 92, 623–638.

Tomz, M. (2007): "Domestic audience costs in international relations: An experimental approach," *International Organization*, 61, 821–840.

Trager, R. F. and L. Vavreck (2011): "The political costs of crisis bargaining: Presidential rhetoric and the role of party," *American Journal of Political Science*, 55, 526–545.

Waterman, R., H. C. Jenkins-Smith, and C. L. Silva (1999): "The Expectations Gap Thesis: Public Attitudes toward an Incumbent President," *Journal of Politics*, 61, 944–966.

Weeks, J. L. (2008): "Autocratic audience costs: Regime type and signaling resolve," *International Organization*, 62, 35–64.

Supplemental Appendix

In the main text, we assumed that the endogenous reference point for voters is updated at two information sets: the one following C's decision not to challenge, and the one following C's decision to challenge. Here, we discuss the equilibrium consequences of alternative modeling choices. We maintain the assumption that in the benchmark model there is no exogenous audience cost, $a = 0$.

The model has three decision points: C's initial decision of whether or not to challenge the territory, D's decision of whether or not to concede, and C's decision of whether to escalate or back down. This means that there are a total of natural possibilities: the reference point is never updated (section A below), the reference point is updated after only the first decision (the case analyzed in the main text), the reference point is update after the first and second decisions (section C below), and the reference point is updated after all three decisions (section B below).

A. The reference point is updated nowhere

If the endogenous reference point is determined (based on rational expectations about equilibrium behavior) at the initial information set and it is *never* updated, then there is no sunk audience cost, $a_s^\rho = 0$. Instead, the tying hands audience cost would be the same as the one we characterized in the main text, $a_t = \beta(1+\eta)z^\lambda$. Then, we can have one of three possible cases:

(i) If $-a_t > z_S > z_W$, then both types of C backs down, D resists and both types of C are indifferent between not challenging and challenging, so each may challenge with any probability.

(ii) If $z_S > z_W > -a_t$, then both types of C choose War, D concedes and both types of C choose to challenge.

(iii) If $z_S > -a_t > z_W$, the strong type of C chooses war, while the weak type chooses to back down. If $q > v/(v + z_D)$ then there is a unique equilibrium in which D concedes and both the strong and weak types of C challenge. If $q < v/(v + z_D)$, then D resists, the strong type of C challenges, and the weak type challenges with any probability weakly larger than $qz_D/(1-q)v$.

B. The reference point is updated everywhere

As mentioned in the main text, if the voters' endogenous reference points are updated at *every* information set of the game, including all terminal information sets, then the equilibrium set of the game is the same as in the augmented model with $\eta = 0$.

Since voters update their reference point at every terminal information set, they cannot be pleasantly surprised or disappointed. As a result, in every equilibrium ρ, $a_s^\rho = 0$ and $a_t = \beta z^\lambda$. Equilibrium behavior is then identical to the one we provided in the main text for the specific case in which $\eta = 0$.

C. The reference point is updated after C's initial choice, and D's choice

Finally, suppose that the endogenous reference point of voters is updated after C's decision of whether or not to challenge, and also after D's decision of whether or not to resist. Then $\mathcal{R}^\rho(\mathcal{I}_d) = 0$, and $\mathcal{R}^\rho(\mathcal{I}_{co}) = v$, where $\mathcal{I}_{co}$ is the information set following D's decision to concede. Also, $\mathcal{R}^\rho(\mathcal{I}_r) = [\tilde{q}\sigma_S^w + (1-\tilde{q})\sigma_W^w]z^\lambda$, where $\mathcal{I}_r$ is the information set following D's decision to resist. The sunk audience cost in any given equilibrium ρ is $a_s^\rho = \beta\eta\mathcal{R}^\rho(\mathcal{I}_r)$, and the tying hands audience cost is again $a_t = \beta(1+\eta)z^\lambda$.

In this case, the equilibria of the game can be pinned down following the same steps we used in the main text. We summarize behavior in the equilibrium set as follows:

(i) If $-a_t > z_S > z_W$, then there is a double continuum of equilibria in which both the strong and weak types of C back down, D resists and each type of C challenges with any probability. Thus, $a_s^\rho = 0$.

(ii) If $z_S > z_W > -a_t$, then there is a unique equilibrium in which both types of C choose war at their final decision nodes, D concedes, and both types of C challenge. In this case, the sunk audience cost is $a_s^\rho = \beta\eta z^\lambda$.

(iii) If $z_S > -a_t > z_W$, then in any equilibrium, the strong type of C chooses war at its final decision node while the weak type backs down. Thus, $a_s^\rho = \beta\eta\tilde{q}z^\lambda$. If $q > v/(v+z_D)$, then both types challenge at the initial decision nodes, D concedes, and $a_s^\rho = \beta\eta q z^\lambda$. Instead, if $q < v/(v+z_D)$, then the strong type challenges at its initial decision node, and we have three subcases:

(a) If $\eta = 0$, then there is a continuum of equilibria in which D resists, and the weak type of C challenges with any probability $\sigma_W \geq qz_D/(1-q)v$. In all of these equilibria, $a_s^\rho = 0$, so there is no sunk audience cost or benefit.

(b) If $\eta > 0$ and $z^\lambda < 0$, there is a unique equilibrium in which D resists and the weak type of C challenges. Thus, there is a sunk audience benefit equal to $a_s^\rho = \beta\eta q z^\lambda < 0$.

(c) If $\eta > 0$ and $z^\lambda > 0$ there is a unique equilibrium in which D resists with probability

$$\sigma_D = \frac{v + z_D}{v + z_D + \beta\eta z^\lambda}$$

and the weak type of C challenges with probability $\sigma_W = q z_D/(1-q)v$. In this case, the sunk audience cost is given by

$$a_s^\rho = \left(\frac{v}{v + z_D}\right) \beta\eta z^\lambda > 0.$$

Thus, behavior in the equilibrium set of the augmented model continues to be analogous to behavior in the equilibrium set of the benchmark model even under the alternative assumption that the endogenous reference points are updated after D's decision as well. It is also straightforward to verify that the comparative statics of the audience costs under this assumption are similar to the comparative statics under the updating assumption made in the main text.