Policy, not Consistency[*]

The distribution of constituency preferences and the true roots of audience costs

Maël van BEEK

2021-Oct-16

Abstract

Students of public opinion and foreign policy know individuals have diverse preferences, yet pragmatism often leads them to abstract these nuances away. Unfortunately, this approach risks obscuring important dynamics. I use audience costs theory – which holds that leaders can tie their hands by issuing public statements since their constituency would punish them for saying one thing but doing another – as case in point. Using a game theoretic model and survey experiments, I propose a theory of audience costs driven by preferences that (1) reproduces the central findings of audience cost theory and reunites it with many of its critiques, (2) reveals how different publics (only hawks, only doves, mixed) shape leader's strategic incentives, (3) sheds light on new dynamics, explaining when leaders are most credible and why they may decide to lie, and that (4) is neither contingent on the presence of a domestic opposition nor vunerable to executive justifications.

[*] I thank Bear Braumoeller, Ryan Brutger, Sarah Croco, Christopher Gelpi, Andrew Goodhardt, Joshua Kertzer, Kathleen McGraw, William Minozzi, David Peterson, Gregory Smith, Daniel Verdier, and Thomas Wood, as well as the participants of the panels at MPSA, ISA and APSA for valuable feedback on earlier drafts of this project. I also thank the Mershon Center for International Security Studies for financial support for this research.

Public opinion is a central feature of our understanding of foreign policy decision-making in democracies. Over the past decades, research has shown leaders know their citizens care about foreign policy, that foreign policy can play a key role during elections, and that angering the public can hurt decision makers at the ballot box or hamstring the rest of their mandate. For some, public opinion is a hindrance to leaders' ability to act with secrecy, flexibility, and expediency; for others, it is a "slumbering giant" that keeps leaders in check (Rosenau 1961) or a double-edged sword that bolsters the credibility of threats but may sometimes compel governments to forsake bargaining altogether (Fearon 1995). It is also a cornerstone of the democratic peace literature where it is used to explain why democracies are less likely to initiate military conflicts – due to the public's aversion to paying the costs of war – and why they may emerge more victorious from international disputes.

Yet, at the heart of this IR scholarship exists a rift: although the behavioral turn and its flourishing literature emphasize that members of the public have diverse preferences, our theories tying public opinion to the foreign policy decision-making process still tend to abstract these nuances away. This disconnect is most visible along methodological lines. Works that employ experimental methods frequently assume individual preferences will eventually influence the foreign policy decision-making process. Conversely, the scholarship that relies on formal or quantitative methods posits the existence of a public available to constrain leaders, often reducing the public to a single unitary actor (e.g., Fearon 1994; Bueno De Mesquita et al. 2005; Schultz 2005). In both cases, these assumptions are driven by pragmatic considerations rather than paradigmatic dogmatism: just as experimentalists know the preferences-to-policy process is far from straightforward, game theorists understand preferences are more diverse than that of the median voter. Rather, abstracting part of the causal chain away allows scholars to focus on their mechanism of interest.

I argue here that, although practical, this approach limits our understanding of how individual preferences inform the foreign policy decision-making process. There are two keys reasons for this. First, we know the public not only has unidirectional likes and dislikes but also multi-directional preferences that vary systematically. Most, if not all, individuals dislike casualties (Mueller 1973) and like victory (Eichenberg 2005) and success (Gelpi, Feaver, and Reifler 2005). Such preferences are unidirectional in the sense they pressure leaders to act in a single way (e.g. to reduce casualties). In contrast, individuals may have different preferences regarding the use of military force (Herrmann, Tetlock, and Visser 1999; Holsti 1979), isolationism (Wittkopf 1990), and free trade (Hainmueller and Hiscox 2006). There is no consensus within the public regarding these preferences which then create multiple contradictory incentives for leaders. Although abstraction can be well suited to explain situations where publics have unidirectional preferences, it cannot account for cases where the public has multi-directional preferences. Leaders held accountable to publics with unidirectional preferences face a straightforward maximization problem. However, leaders who face publics with multi-directional preferences on substantive issues cannot always satisfy the public as a whole. Often, pleasing one group means disappointing another and these leaders receive a flat reward or cost for adopting a policy, contingent on its substance. As a result, this tradeoff can incentivize leaders to adopt policies for reasons orthogonal to their inherent quality and to pander to the public in the hope of a reward – or at least to avoid punishment.

Second, we also know that the public is not static. Changes in the public can be the result of evolving individual preferences, for example, in response to world events (Holsti 2004); amendments to domestic institutions, such as when the franchise was expanded (Barnhart et al. 2020); or shifts in the internal ordering of the electorate, as in the case of the sorting of individuals into more homogeneous parties (Levendusky 2009). Regardless of the reason, as the public's preferences and composition mutate over time, so do the incentives and constrains leaders must face. Ignoring these dynamics means we risk misconstruing the strategic environment that shapes leaders' decision-making process as well as overlooking important facets of international politics.

To demonstrate the importance of addressing this divide, I focus on a prevalent model tying public opinion to foreign policy outcomes in IR: audience costs – the argument that leaders can tie their own hands by making declarations since their domestic public will punish governments who say one thing but do another (Fearon 1994; Baum 2004; Slantchev 2006; Weeks 2008). Initially used to explain military threats, this literature frequently brackets one of the best established divides in the IR public opinion research: that between pro-war hawks and pro-peace doves which has been repeatedly associated with support for U.S. military involvement abroad (Hurwitz and Peffley 1987; Holsti and Rosenau 1993; Herrmann, Tetlock, and Visser 1999).

I account for this split by expanding on the classic crisis-bargaining model and presenting a game-theoretic model where the leader's payoffs are a function of the distribution of preferences regarding the use of force within her public. I then turn to an experimental design to provide direct evidence of these costs at the individual level. This structure mirrors the history of the audience costs literature where audience costs were first introduced as part of a formal model (Fearon 1994) before their existence was tested using survey experiments (Tomz 2007).

My results highlight the importance of preserving the causal chain linking public opinion to the foreign policy decision-making process. Ironically, I find that assuming the public is homogeneous obscures important strategic dynamics in both mixed and homogeneous publics. In the context of audience costs, I present four key findings. First, inconsistent leaders are punished because members of the public have substantive policy preferences – not because they dislike inconsistency. This is an important departure from the literature on audience costs that assumes these costs do "not arise because domestic audiences disagree with the leader's policy" (Weeks 2008, 43) but because the public dislikes inconsistency, an assumption corroborated in many experiments (Davies and Johns 2013; Levy et al. 2015; Quek 2017; Tomz 2007). Surprisingly, this preferences mechanism is the same as that used by critics to challenge the existence of audience costs (Chaudoin 2014; Snyder and Borghard 2011; Trachtenberg 2012). In fact, it is because individuals have preferences over the substance of policies that previous experimental studies were able to observe the dynamics predicted by audience costs theory. Because these studies relied on samples that included both hawks and doves, the punishment enforced by one group was offset by the reward granted by the other, leaving only the cost of betrayal that ties thes hands of leaders to be observed. Second, the composition of the public generates drastically different incentives for leaders. Publics with mixed preferences (that include both hawks and doves) produce no incentive for leaders to follow a specific policy (only to remain consistent) and allow for the creation of modest hand-tying costs for whichever policy the leader selects. Publics with homogeneous preferences (with only hawks or only doves), however, create important incentives in favor of their preferred policy while generating strong hand-tying costs against reneging on this policy. Third, these dynamics carry substantial implications for crisis bargaining, from explaining when leaders are most credible and why they may decide to lie, to showing how the substance of policies impacts how conspicuous a leader's inconsistency is likely to be. Fourth, these dynamics are robust to the common belief that the enforcement of audience costs is contingent on a domestic opposition willing to criticize inconsistent leaders as well as to the executive spin critique (Levendusky and Horowitz 2012). Using an original survey design, I find evidence that individuals will punish leaders who betrayed them even if the opposition remains silent about the leader's inconsistency. I also find individuals do not forgive presidents who claim to have received new information; instead, they simultaneously reward presidents for their diligence and reward (or punish) them for the susbstance of their policy.

1 Preferences and Audience Costs

Simply defined, audience costs are the penalty state leaders incur when they back down from a publicly stated commitments – it is the cost leaders have to pay for saying one thing and doing another. Initially used in the context of a leader backing down from threats during a military crisis, audience costs have been applied to a variety of situations, including promises to allies (Tomz 2007), international cooperation (Leeds 1999), the impact of central bank on inflation (Broz 2002), and economic disputes (Busch 2000). Although the specific mechanism at work has yet to be specified (Croco, Hanmer, and McDonald 2020; Kurizaki and Whang 2015, 950), the existence of audience costs has been premised on the working assumption that the public is willing to punish the inconsistency of leaders. Different justifications have been given for this assumption: for example, the public could want to remove from office leaders who have tarnished the reputation of the nation or revealed their incompetence in foreign affairs (Fearon 1994; Guisinger and Smith 2002; Schultz 2001; Smith 1998). Regardless of the specific reason, however, the theoretical expectation remains the same: the public de facto exhibits an homogeneous aversion to inconsistency.[1]

Over the past decade, however, Traditional Audience Costs Theory (TACT) has come under repeated attack (Downes and Sechser 2012; Mercer 2012).[2] These attacks fall under

[1] Nomikos and Sambanis (2019) find that incompetence is an important component of the punishment audiences enforce on leaders. Despite using similar language, this project investigates a different process. Nomikos and Sambanis zoom in on the ability of leaders to deliver succesful outcomes in terms of policy success or failure (whether the military action succeeded or failed at pushing back the attacking state). Consistent with the literature, they find that leaders who are defeated on the battlefield are punished compared to victorious ones. In contrast, I hold policy success constant and focus instead on the substance of the policy delivered (whether the president eventually ordered intervention or not).

[2] Some works have come in defense of TACT, e.g. Baum and Potter (2015) argue the type of democracy

three categories. The first critique challenges the validity of the inconsistency assumption by asking why a rational public would want to punish a rational leader. Some scholars argue that, because individuals have substantive preferences regarding the content of policies, the public has no interest in imposing costs on inconsistent leaders. After reviewing prominent international crises, Snyder and Borghard (2011) and Trachtenberg (2012) conclude domestic audience costs have a minimal effect on crisis behavior because "domestic audiences understandably care more about policy substance than about consistency between a leader's words and deeds" (Snyder and Borghard 2011, 55) and backing off "might not have major political consequences at home, especially if the 'audience' wants to avoid war" (Trachtenberg 2012, 39). Consistent with this view, Chaudoin (2014) finds, using survey experiments, that only respondents with no opinion on trade policy punish inconsistent leaders. In contrast, respondents with strong preferences over free trade have muted reactions to learning that their leader has broken an agreement.[3] The second critique focuses on the electoral mechanism implied by TACT and contends audience costs are contingent on the existence of a domestic opponent willing to criticize the leader for her inconsistency or that presidential approval, the DV most commonly used in audience costs experiments, does not capture this electoral mechanism accurately (Croco, Hanmer, and McDonald 2020). The third critique argues executive leaders can shirk punishment by announcing they have received new information (Levendusky and Horowitz 2012). Here, neither inconsistency nor preferences matter as the leader simply spins the news in her favor. These attacks have important implications: as Mercer (2012, 399) notes, "if audience cost mechanisms are imaginary, then so are the solutions that rely on them."

Two characteristics of the audience costs literature make it a challenging candidate for testing the distribution of preferences proposition. First, audience costs proponents and opponents alike have adopted a unitary approach to the public and its preferences. Regardless of whether it dislikes inconsistency itself (Fearon 1994; Weeks 2008); opposes military intervention (Snyder and Borghard 2011); or is susceptible to new information (Levendusky and Horowitz 2012); the public is assumed to have a unidirectional preference. Second, despite support for the use of military force being a well established divide in the public (Holsti 2004; Hurwitz and Peffley 1987; Kertzer et al. 2014; Rathbun et al. 2016), most survey experiments have ignored the role of hawks and doves and found repeated support for the claim that the public imposes audience costs on inconsistent leaders (Davies and Johns 2013; Levendusky and Horowitz 2012; Levy et al. 2015; Quek

matters and Trager and Vavreck (2011) highlight threat specificity. I argue here that, beyond these external defenses, there are also internal reasons to believe audience costs matter.

[3] This work differs from Chaudoin (2014) in three ways. First, I show that preferences matter at the individual level by expanding on Tomz (2007)'s experimental design which has repeatedly produced results in support the consitency position. Second, whereas Chaudoin (2014) finds that only individuals with weak preferences punish inconsistent leaders, I show that both hawks and doves are willing to punish a specific kind of inconsistency (betrayal). Third, I expand on the two projects' shared intuition by investigating the strategic implications of diverging public preferences for international bargaining. This approach reveals that the political calculus of leaders who have made public commitments is in fact different from that of leaders who have not (c.f. Chaudoin 2014, 253). Moreover, I demonstrate that publics with only stong hawks and doves (and no individual with weak preferences) produce *ex ante* strategic environments for leaders identical to those predicted by TACT.

2017; Tomz 2007; Trager and Vavreck 2011).

Focusing on the link between divided public and the executive decision-making process reunites audience costs with its critiques and offers a precise mechanism, public preferences, linking leaders' behavior to punishment.

This project expands upon a growing literature that seeks to endogenize public preferences regarding foreign policy into models of public attitudes and foreign policy.[4] Most salient, in the audience costs literature, is the work of Kertzer and Brutger (2016) which proposes that audiences can punish leaders for being inconsistent or for their belligerence. That is for threatening the use of force in the first place. Additionally, they find that, when the president emits a threat but eventually backs down, individuals with hawkish preferences punish inconsistency, while individuals with dovish preferences punish belligerence. Unfortunately, the relation between individual preferences and the substance of policies make these findings difficult to generalize. This is apparent when focusing on the reverse scenario: if the president backs in, by initially announcing the army will stay out of the conflict but eventually ordering the army to intervene anyway,[5] should hawks still be expected to punish inconsistency, and doves belligerence?

Here, I propose a general theory of public preferences and audience costs, the Preferences Model of Audience Costs (PACT), that incorporates not only the specific cases identified by Kertzer and Brutger (2016)[6] but also TACT. It is important to notice that "Backing down" has different substantive implications depending on whether respondents have hawkish or dovish preferences. For hawks, the president has betrayed them by failing to deliver a desired policy; for doves, however, the president has seen the light by abandoning the unpopular policy she had initially adopted. Through this lens, it makes intuitive sense that hawks would punish inconsistency and doves would punish belligerence. In contrast, when the president "Backs In," the situation is reversed: hawks are vindicated and doves betrayed. Since Backing In is the mirror image of Backing Down, I expect the costs enforced by the audience are also mirrored. Thus, I argue that if the president backs in, hawks would punish the leader's initial pacifism, while doves would punish the leader for being inconsistent and betraying them.

Conceptually, I argue that the belligerence cost enforced by doves is an example of broader preference costs which also include a pacifism cost enforced by hawks. Subordinate to these preferences costs are two kinds of inconsistency costs: betrayal costs, the costs enforced when leaders abandon a liked policy, and redemption costs, the costs of

[4] I focus here on the distribution of preferences in the public. Others have focused on the preferences of leaders: i.a. Kreps, Saunders, and Schultz (2018); Mattes and Weeks (2019); Heffington (2016).

[5] Although the audience costs literature has mostly focused on Backing Out, scholars have found the public is also willing to enforce costs for leaders who Back In (Levy et al. 2015; Quek 2017). Both costs share the same logic: the public punishes the leader for saying one thing and doing another. This is consistent with the traditional dislike for inconsistency assumed by TACT. Unlike TACT, the theory outlined hereafter can account for the variation between the two costs.

[6] Compared to Kertzer and Brutger, 2016, this work shows that the inconsistency costs they observed are predicted by individual preferences. Furthermore, this project also models the implications of the distribution of preferences on the incentives of leaders and addresses the new information critique.

abandoning an unpopular policy. Put simply, all individuals like leaders who deliver substantive policies they approve of, dislike those who do not, and hate being betrayed.

This generalized theory also subsumes TACT as a special cases of PACT only obtained when the public has mixed preferences (when it includes both hawks and doves). In such a situation, the preference costs and rewards of both sides cancel out and only the cost of betrayal remains to punish a leader's inconsistency.

In the following sections, I present PACT as game theoretical model to demonstrate it reproduces the findings of TACT and reveals new dynamics between public opinion and leaders. Then, I use original survey experiments to show the public does enforce these costs and address the electoral and executive spin critiques.

2 The Preference Model of Audience Costs Theory

The Preference model of Audience Costs Theory (PACT) expands on the classic crisis-bargaining game common in this literature from which it differs in two ways: first, the president's payoff are a function of the distribution of preferences within her constituency; second, the president has the opportunity to intervene against the challenger after announcing she would stay out of the conflict. Though used here in the context of a military crisis, the game's insights can also be applied to public statements during bargains more generally.

It contains two strategic players, a leader (S_1) and a foreign challenger (S_2), and starts with the leader declaring her intent to intervene (*threat*) or stay out (*assurances*) of a conflict initiated by the foreign challenger. The foreign challenger then decides whether or not to persevere in his aggression. If he stand downs, the game ends. If he continues, the leader must decide whether to follow through and be consistent, or to abandon the policy thereby becoming inconsistent. This makes the substantive policy selected contingent on which declaration the leader previously issued: sending troops and engaging the foreign armies is consistent only if she declared initially she would intervene – and vice versa.

For state i, the payoffs consist of the value of the disputed issue, $v_i > 0$, and of the cost of military conflict, $c_i > 0$, when applicable.

Central to this model is the fact that the leader's (S_1) payoffs also include costs derived from the audience. Let the interval $C = [0, 1]$ be the leader's constituency space, with each point $\alpha \in C$ representing a particular audience with a specific proportion of hawks, α, and doves, $1 - \alpha$. Let $p_j \geq 0$ be the strength of the preferences held by domestic group j. p_j is added to the leader's payoff when the declared policy is consonnant with the j's preferences and subtracted when it is not. Last, let $b_j \geq 0$ be the punishment j inflicts on the leader when betrayed.

All payoffs consist of an audience component and, when applicable, of a prize and cost of war component. The audience component is contingent on the leader's choices at both nodes. At the first node, threats are rewarded by hawks, αp_h, but punished by doves, $(1 - \alpha)p_d$ – and vice-versa in the absence of threats. At the second node, if she is consistent, the leader's payoffs remain unchanged. However, if she contradicts herself, she must pay a betrayal cost enforced by the group she disappointed: αb_h enforced by hawks if she emitted a threat; and $(1-\alpha)b_d$ enforced by doves if she had not. Whenever the challenger complies, either by conceding or as the result of a military intervention by the leader, the leader adds the value of the disputed issue, v_1, to her payoff. Finally, whenever she intervenes, the leader must substract the cost of intervention, c_1, from her payoff.

To give a specific example, a leader who threatened intervention and is facing a foreign challenger who refused to concede will be consistent (and intervene) only if

$$\alpha p_h - (1-\alpha)p_d + v_1 - c_1 \geq \alpha p_h - (1-\alpha)p_d - \alpha b_h \tag{1}$$

For simplicity, I assume hawks and doves reward and punish the leader at the same rate, standardized to 1 ($p_h = p_d = b_h = b_d = 1$). Equation 1 thus simplifies to

$$2\alpha - 1 + v_1 - c_1 \geq \alpha - 1 \tag{2}$$

where $2\alpha - 1$ and $a - 1$ represent the audience components and $v_1 - c_1$ the prize and war components. Thus, in this scenario, the leader will be consistent and intervene only if $v_1 - c_1 \geq -\alpha$; if the total value of intervention is greater than the value of betraying hawks. Figure 1 shows the game's simplified extended form.

The solution concept is the subgame perfect Nash equilibrium and is provided in the appendix. Figure 2 shows State 1's equilibrium strategy for $v_1 = 1$.

3 Preferences Explain Audience Costs and More

The discussion of this model proceeds in three steps. First, I reproduce the traditional audience costs equilibria to demonstrate that individuals having preferences over policy substance is compatible with the existence of audience costs. Second, I compare constituencies with heterogeneous preferences to those with homogeneous ones and find that audiences can create incentives for leaders to be inconsistent. Third, I show that whether the audience is homogeneously hawkish or dovish carries important implications for leaders' behavior.

Under TACT, a leader will be forced to pay costs if she says one thing but does another. This could happen because she threatened intervention but eventually backed down from the conflict and, as a result, performed worse than if she had stayed out of the conflict entirely. Alternatively, this could also happen because she announced her intention to stay out of the conflict but eventually backed in and, consequently, performed worse than if she had declared her intention to intervene immediately. In PACT, the former amounts to positing: $\alpha - 1 \leq 1 - 2\alpha$; the latter: $2\alpha - 1 + v_1 - c_1 \geq -\alpha + v_1 - c_1$. Both inequalities are satisfied for $1/3 \leq \alpha \leq 2/3$, that is when the public is heterogeneous and includes

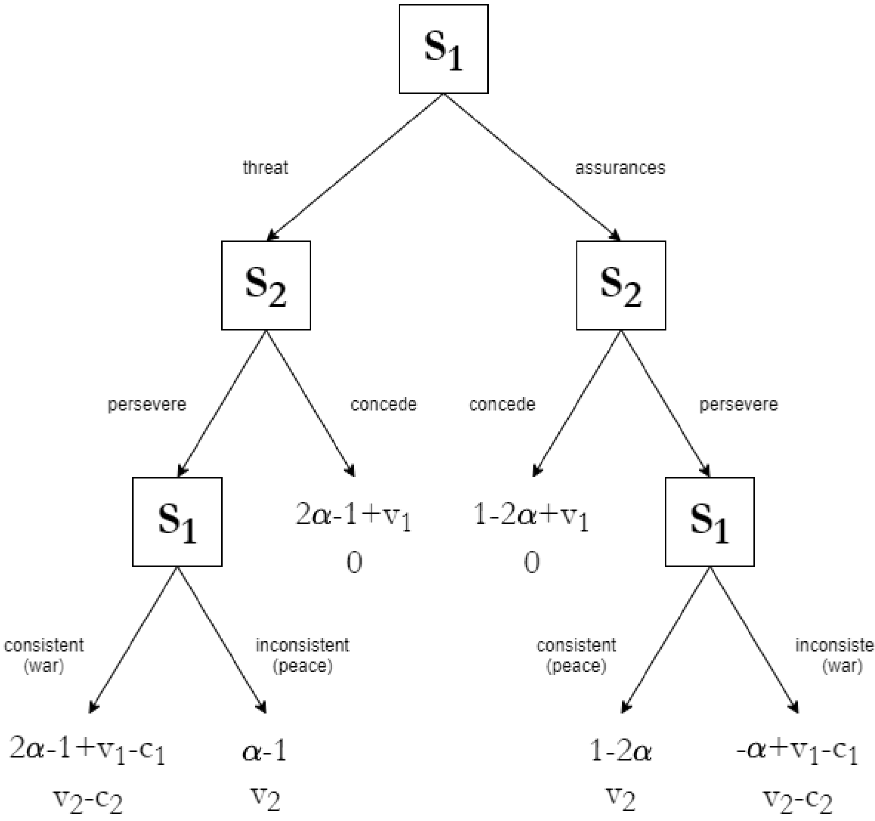

Figure 1: Game tree with simplified payoffs (S_1 on the first line; S_2's on the second)

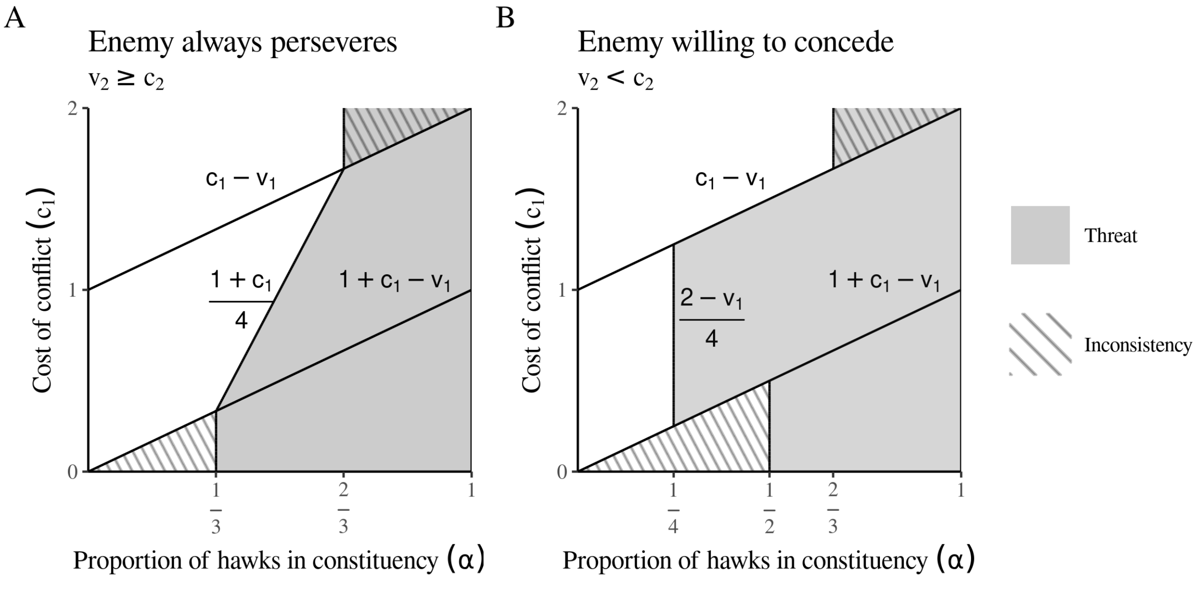

Figure 2: S_1's equilibrium strategy ($v_1 = 1$)

both hawks and doves. When accountable to such an audience, leaders are always able to produce hand-tying costs – as argued by TACT.

This finding hinges on the fact that, in the case of heterogeneous constituencies, the punishment the leader suffers for announcing any given policy is countered by a similarly sized reward. However, the cost of betrayal (enforced by the members of the constituency in favor of the policy) has no counterweight and the leader is thus always better off following through on her word. For example, if the leader of a heterogeneous constituency adopts a hawkish policy, she will be rewarded by hawks and punished by doves. In practice, since the two groups are similarly sized, the rewards granted by hawks and the costs enforced by doves for announcing this policy cancel out.[7] Once the policy is adopted, the leader has incentives to be consistent as hawks would punish her for betraying them by ordering the troops to stand down but doves would not reward her redemption. If the leader had instead announced she would stay out of the conflict, the roles of hawks and doves would be reversed. In this case, rewards and costs for announcing a dovish policy would balance out once again and the leader would have incentives to remain true to her words as intervention would be understood as a betrayal by doves who would then punish her. The central TACT argument – that audiences punish leaders who say one thing but do another – is thus reproduced by PACT.

The fact that preferences reproduce the traditional dynamics of audience costs has important implications. Theoretically, the substantive policy position is not the challenge to audience costs some critics believe it to be. Methodologically, if PACT and TACT predict identical outcomes, more attention needs to be paid to the specific conditions under which the two competing models can be disentangled. Doing so requires turning to the extremes: when publics are not heterogeneous but homogeneous and include only hawks or only doves.

Unlike heterogeneous constituencies which encourage consistency, homogeneous constituencies (when $\alpha < 1/3$ or $\alpha > 2/3$) can create incentives for their leaders to be inconsistent and lie. Even though the leader knows intervention is not an option (specifically, $\alpha < c_1 - v_1$), she will still issue a public threat if her audience is strongly hawkish ($\alpha > 2/3$). Similarly, a leader who knows intervention is forthcoming ($1 - v_1 + c_1 < \alpha$) will still pretend to have no intention of being involved in the conflict when she is facing a strongly dovish public ($\alpha < 1/3$). Both leaders know their inconsistency will be exposed, yet they still prefer lying to adopting an unpopular position. Why is that?

Consider the case where a leader is in the midst of a crisis where intervention yields only limited benefits but is prohibitively costly (e.g., if the disputed issue has minimal strategic value and the aggressor has nuclear capabilities). Because intervention is not an option, the leader must choose between truthfully declaring her intention of staying out of the conflict or misrepresenting her intentions by issuing a threat before eventually backing down. For TACT, this case is evident: the leader will always be truthful to avoid being

[7] Technically, rewards and costs derived from the audience only cancel out if $\alpha = 0.5$. Nonetheless, for the range $1/3 \leq \alpha \leq 2/3$, the marginal utility of adopting any given policy is substantively insignificant and remains secondary to other strategic considerations.

punished for her inconsistency. This is not true for PACT. Imagine now that this leader must answer to an audience that only includes individuals with hawkish preferences. If she is honest, her audience will punish her for adopting a dovish policy. In contrast, if she lies, she will first be rewarded for her initial hawkishness, and later punished for her later betrayal. In this situation, although the initial reward is eventually canceled by this punishment, lying remains a more attractive option than honesty as honesty is immediately punished but never rewarded.[8] This is fundamentally different from the predictions made by TACT: not only does a domestic audience not deter leaders from inconsistency but it can actually incentivize them to be inconsistent.

To be noted, the propensity of leaders of homogeneous publics to belie their intentions does not necessarily imply a credibility deficit. Rather, when leaders face homogeneous publics, the credibility of any policy is contingent on its consonance with public preferences. Popular policies are highly credible as the entire constituency would be willing to punish inconsistency – and leaders thus enjoy strong hand-tying costs. Unpopular policies, however, enjoy no such benefit as no member of the public would be disappointed by a sudden change in policy. Taken together, these dynamics suggest the existence of a tradeoff between leaders' agency (in terms of policy choice) and the credibility of their policies.[9]

Finally, I turn to the strategic contexts generated by different audiences. The leader of a hawkish audience ($\alpha > 2/3$) has a dominant strategy at the level of the first node: to always threaten. She will do so regardless of whether the challenger is expected to persevere or concede and regardless of whether she intends to intervene or not. In contrast, the strategy of a leader answering to a dovish audience ($\alpha < 1/3$) depends on the expected choice of her challenger. If persevering is the challenger's dominant strategy ($v_2 > c_2$), then she will never threaten, regardless of whether she intends to intervene or not. However, if the challenger is willing to concede when presented with a credible threat ($v_2 < c_2$), the situation is more complex and the challenger's strategic considerations can induce the leader to emit a threat.[10] This is a function of the overall utility of intervention for the leader. When intervention is so prohibitively costly that it is never an option ($\alpha < c_1 - v_1$), or so worthwhile that staying out is never credible ($\alpha > 1 + c_1 - v_1$), the leader will not issue a threat. However, in the middle range ($c_1 - v_1 < \alpha < 1 + c_1 - v_1$), the leader of a dovish constituency may be willing to defy popular preferences if doing so induces the challenger to comply. This result obtains whenever the value of the disputed issue outweighs the cost of defying public preferences ($\alpha > \frac{2-v_1}{4}$).

This nuance between the strategies induced by each kind of audience is far from trivial.

[8] In the mirror scenario, when the utility of intervention is so great military engagement is certain, a dovish audience can create incentives for leaders to lie and state they will not intervene rather than reveal about their bellicose intentions.

[9] Additionally, homogeneous publics provide strong incentives for leaders to adopt their favored policy and avoid disliked policy.

[10] When the challenger is willing to concede, the threshold for an audience to be considered "dovish" rises to $\alpha < 1/2$.

PACT suggests that the leaders of hawkish audiences should be observed being inconsistent more frequently than those of dovish audiences. Consider a challenger who would be willing to concede when presented with a credible threat ($v_2 < c_2$). When facing the leader of a hawkish constituency who issues a threat she has no intention of enforcing, the challenger's best response is to persevere. However, when the challenger faces the leader of a dovish constituency who lies by indicating that she intends to stay out of the conflict but will surely intervene, he knows his best response is to concede preemptively. Yet, because this concession occurs prior to intervention taking place, the leader's inconsistency is not revealed and remains unobserved. As a result, the inconsistency of dovish leaders should be expected to be less visible than that of hawkish leaders, not because they are less prone to inconsistency but because their lies are not as likely to be exposed.

A counterintuitive implication of PACT is that a more hawkish constituency does not always increase the likelihood of a leader emitting a threat. This is visible in Figure 2*B* where, for $\frac{1}{4} < c_1 < \frac{1}{2}$, the proportion of hawks in the public has a non-linear relationship with the likelihood of issuing a threat. When the audience includes few hawks, the leader is best served by expressing her truthful intention to stay out of the conflict as a threat would not be credible and she would be punished for making one. As the proportion of hawks increases, however, it becomes worthwhile (starting at $\frac{2-v_1}{4}$) to issue a threat as doing so ensures a concession from the challenger. This is because there are just enough hawks in the public to make the threat of intervention credible. As the number of hawks increases still, it reaches a threshold ($\alpha > 1 + c_1 - v_1$) where intervention becomes certain and the challenger would concede even in the absence of a threat. Exploiting this opportunity, the leader can now refrain from issuing a threat and please the doves that are still the majority of her constituency. Finally, as hawks become the majority of the audience ($\alpha > \frac{1}{2}$), the leader now caters to hawkish preferences and issues a threat. By symmetry, this also means a more dovish public may cause leaders to emit threats they would not have issued otherwise.

In sum, focusing on the distribution of preferences delivers three new and counterintuitive results: First, PACT reproduces all the results of TACT in publics with heterogeneous preferences. This shows that members of the public having preferences over the substance of policies does not pose a challenge to the scholarship that relies on the assumption that leaders are be punished for saying one thing but doing another. Second, costs derived from audiences may incentivize leaders to be – and not detract them from being – inconsistent on the international stage. Third, audiences with different substantive preferences (hawkish or dovish) generate different costs and strategic environments for leaders.

To confirm that publics do enforce these costs, in the next section I turn to a survey experiment. Consistent with my model's predictions, I find that: first, mixed publics punish inconsistent leaders; second, mixed publics do not punish or reward the susbstance of policies; third, that homogeneous publics punish leaders who betray them; fourth, that homogeneous publics do not punish leaders who "see the light"; and fifth, that homogeneous publics do punish and reward leaders for the substance of the policies they adopt.

4 Testing PACT

The previous section has established that traditional audience costs can be the product of public preferences over the substance of policies, that these preferences can incentivize leaders to be inconsistent, and that whether the public is dovish or hawkish matters. I use a series of survey experiments to provide empirical evidence for PACT's propositions.

4.1 Experimental Design

An online survey experiment was fielded in August 2020 on a national American sample of 1130 citizens of voting age recruited by Lucid. The survey builds on Tomz (2007)'s canonical setting and its expansion by Kertzer and Brutger (2016). Participants are presented with a short text explaining they will read about a situation the U.S. has faced in the past and may face again; that different leaders have handled the situation differently; and that they will be asked whether they approve or not of the president's actions. They are then told an unidentified foreign country has sent its military to invade a neighboring nation. Participants are then randomly assigned to two treatments:

> [`threat`] The U.S. president announced that if the attacking country continued to invade, the U.S. military would immediately engage and attempt to push out the attacking country.
>
> [`assurance`] The U.S. president announced that even if the attacking country continued to invade, the U.S. military would stay out of the conflict.
>
> Following the presidential statement, the attacking country continued to invade its neighbor.
>
> [`consistency / inconsistency`] The U.S. president then ordered the troops [`to engage` / `not to engage`] the attacking country.
>
> In the end, no U.S. solider has died during this conflict and the attacking country has gained control of 20% of the contested territory.

Finally, three dependent variables were measured. The first is the traditional approval question where participants are asked the extent to which they approve or disapprove of the way the president handled the situation, on a seven-point scale. The other two are novel measures meant to provide a more finely-tuned test of the electoral mechanism implied by the audience costs literature. These were presented in sequence:

> The U.S. president in the text you just read about is running for reelection and is facing a challenger during the primaries.
>
> Being from the same party, both candidates have adopted similar positions on domestic matters.
>
> The challenger is a vocal opponent of the president's foreign policy and argues that the U.S. military should have [`stayed out of`/`intervened in`] the conflict.

> If the election were today, whom would you vote for?

Participants could then choose between the president and the challenger. After their decision, they were asked to imagine another kind of challenger:

> Now imagine instead that the challenger has remained silent on the issue of intervention but publicly opposes the president's foreign economic policy, arguing that the trade policies undertaken by the president's administration were a mistake.
>
> If the election were today, whom would you vote for?

To avoid ordering effects, these challengers were introduced in a random order.

Individual preferences, how hawkish or dovish respondents are, were captured using militant assertiveness items borrowed from Herrmann, Tetlock, and Visser (1999) also used by Kertzer and Brutger (2016). Following the approach of Kertzer and Brutger (2016), I separate participants along their militant assertiveness score to construct these publics: the hawkish sample includes participants who placed in the top 25%; and the dovish sample, those who in the bottom 25%. In addition to these two homogeneous publics, I also discuss two mixed publics: the full sample and a balanced sample. As in previous studies, I test my hypotheses against the full sample. Although this sample is mixed and includes both hawks and doves, these groups are not present in equal numbers and it leans towards the hawkish position.[11] As robustness check, I construct a balanced sample by randomly selecting an equal number of hawks and doves. Party affiliation and standard demographic characteristics, including gender and ethnicity were also measured. Treatment wordings can be found in the appendix.

4.2 Identifying Effects of Interest

The design described above has four experimental conditions which reflect the four final outcomes of the game visible at the bottom of Figure 1. These are: Threat & Consistency (`T&C`), Threat & Inconsistency (`T&I`), Assurance & Consistency (`A&C`), and Assurance & Inconsistency (`A&I`).

I use these experimental conditions to identify the effect of four primary theoretical variables: two kinds of preferences (Consonance and Dissonance) and inconsistency (Betrayal and Redemption) effects.

The effect of Consonance is the effect of adopting and delivering a popular policy compared to an unpopular one (for hawks, `T&C`−`A&C`; for doves, `A&C`−`T&C`) while the effect of Dissonance is the cost the leader must pay for doing the opposite (for hawks, `A&C`−`T&C`; for doves, `T&C`−`A&C`). While Consonance and Dissonance are substantively distinct concepts, this distinction is mathematically trivial as the former is the additive inverse of the latter ($Consonance = -Dissonance$). For the sake of clarity, I only discuss the effect of

[11] The average militant assertiveness of this sample is 2.1, where 0 is the least hawkish (most dovish) position and 4 the most hawkish (least dovish) position. For comparison, the dovish and hawkish samples score 1.07 and 2.98, respectively, on the same scale.

Consonance. Finally, the effect of Betrayal is the effect of abandoning a policy the public likes instead of following through on it (for hawks, T&I−T&C; for doves, A&I−A&C) and the effect of Redemption is the effect of abandoning a policy disliked by the public instead of following through on it (for hawks, A&I−A&C; for doves, T&I−T&C). Table 1 recapitulates these effects.

Table 1: Effects in formal and experimental notations

Effect	Formal	For hawks	For doves
Consonance	$(p_j) - (-p_j)$	T&C−A&C	A&C−T&C
Dissonance	$(-p_j) - p_j$	A&C−T&C	T&C−A&C
Betrayal	$(p_j - b_j) - p_j$	T&I−T&C	A&I−A&C
Redemption	$(-p_j) - (-p_j)$	A&I−A&C	T&I−T&C

4.3 Experimental Hypotheses

I use this experimental design to test PACT and provide supporting evidence that different publics enforce different costs based on their preferences. I first present the hypotheses for the mixed publics before discussing those for the hawkish and dovish publics.

Since TACT can be understood as a special case of PACT that holds only in mixed publics, both TACT and PACT have the same expectations regarding the dynamics exhibited in mixed publics. Since these publics include both hawks and doves, PACT would expect that the dissonance costs imposed by one group would be alleviated by the consonance rewards offered by the other group. In contrast, the betrayal costs would not be alleviated by redemption rewards and the costs of inconsistency would be observable. The language of betrayal, redemption and consonance is ill adapted to discussing mixed public due to the presence of conflicting preferences. Intead, I fall back on the broader language of inconsistency and preferences costs.

- H_1: In mixed publics, the effect of Inconsistency is negative.
  - H_{1A} After a threat: T&I−T&C < 0
  - H_{1B} Without a threat: A&I−A&C < 0
- H_2: In mixed publics, the effect of Preferences is *not* negative.
  - H_{2A} Belligerence: T&C−A&C ≥ 0
  - H_{2B} Pacifism: A&C−T&C ≥ 0

In homogeneous public, TACT and PACT diverge. Implicticly, the expection for TACT would be that homogeneous publics would behave like mixed publics. In contrast, PACT predicts that homogeneous publics will only punish leaders who betray them (not those who have redeemed themselves), and that they will reward leaders who adopt the policies they like. This amounts to stating that:

- H_3: In homogeneous publics, the effect of Betrayal is negative.
  - For hawks: T&I−T&C < 0
  - For doves: A&I−A&C < 0

- H_4: In homogeneous publics, the effect of Redemption is *not* negative.
  - For hawks: $\texttt{A\&I} - \texttt{A\&C} \geq 0$
  - For doves: $\texttt{T\&I} - \texttt{T\&C} \geq 0$
- H_5: In homogeneous publics, the effect of Consonance is positive.
  - For hawks: $\texttt{T\&C} - \texttt{A\&C} > 0$
  - For doves: $\texttt{A\&C} - \texttt{T\&C} > 0$

5 Constituencies generate costs consistent with PACT

The experimental results provide strong support for PACT. Focusing on different publics reveals that individual preferences drive the costs leaders face. Leaders are rewarded for adopting popular policies, penalized for endorsing unpopular ones, and only punished for their inconsistency if it yields an outcome the public dislikes.

The average treatment effects and the probability distributions for the values presented in this section were derived from $2,000$ nonparametric bootstraps. This approach is conservative and makes no assumption regarding the distribution of the population. Due to the strong theoretical priors regarding the direction of these effects, the p values presented here are for one-tailed tests with 95% confidence intervals.[12]

Key to this approach is the randomization of treatment assignments which, on average, makes the different groups equal in terms of all characteristics.[13]

5.1 Replicating TACT

Because the design of this study and the sample used are comparable to those of prior studies, I expect that the traditional experimental findings in favor of TACT should be replicated here as well. I show that this is indeed the case. It is the combination of having all the possible outcomes as well as capturing participants' preferences that make it possible to validate PACT's hypotheses in the different samples.[14] Before that, however, I focus on the mixed publics and specifically the full sample.

As expected, the canonical findings of TACT hold true: Backing Out and Backing In are both costly. When the leader threatens engagement but eventually backs down, compared to if she had stayed out of the conflict in the first place ($\texttt{T\&I} - \texttt{A\&C}$), she suffers a loss of 0.311 point ($p < 0.032$) in approval rating on a seven-point scale.[15]

[12] I use the percentile method (at the 95th percentile) for the one-tailed p values. I use the bias-corrected and accelerated (BCa) method to identify the 95% confidence intervals of the two-tailed test. Unless stated otherwise, all the effects significant in a one-tailed test are also significant in two-tailed tests.

[13] As visible in the appendix, I find no evidence treatments or treatment combinations were assigned non-randomly with respect to various pretreatment variables.

[14] The designs of Levy et al. (2015) and Quek (2017) include all the possible outcomes, and Kertzer and Brutger (2016) capture participants' preferences but this is, to our knowledge, the first study that combines these features.

[15] This result is at the margin of the traditional statistical significance threshold for a two-tailed test (-0.664, 0.028).

Similarly, when she announces the army will stay out of the conflict but eventually backs in, she fares worse than if she had intervened immediately. On average, when Backing In (A&I−T&C), her approval rating is 0.582 point lower ($p < 0.001$). These results replicate the findings of Levy et al. (2015) and Quek (2017).

As Kertzer and Brutger (2016) note, however, this setup is double barreled. I decompose audience costs into its two components: the inconsistency cost leaders have to pay for abandoning a policy; and the preferences cost they must pay for defying the will of the people.

Leaders can suffer two sorts of inconsistency costs: one after emitting a threat (H_{1A}), the other after declaring the state will stay out of the conflict (H_{1B}). TACT predicts inconsistency, no matter its substance, should be punished. Figure 3 shows the combined effect of these two kinds of inconsistency. Consistent with the expectations of TACT, inconsistency is severely punished. Overall, the average treatment effect of inconsistency on approval ratings is -0.447 ($p < 0.001$).[16]

Finally, I also find no evidence that the effect of preferences is negative. Leaders who emit a threat and follow through on it score on average 0.312 points *higher* than those who promise and deliver peace (T&C−A&C). This effect is significantly different from 0 when using a one-tailed test ($p < 0.04$) and at the margin of the traditional statistical significance threshold using a two-tailed test (-0.026, 0.65). These results are presented in Figure 3. This is a suprising departure from Kertzer and Brutger (2016) who find that the public enforces a statistically significant cost on belligerent leaders, not rewards them.

Overall, these findings confirm TACT and replicate the traditional findings of previous audience costs experiments. Figure 3 also presents the results for the balanced public which displays the same dynamics as the full sample. Saliently, since TACT is a special case of PACT for mixed publics, these results provide support for PACT.

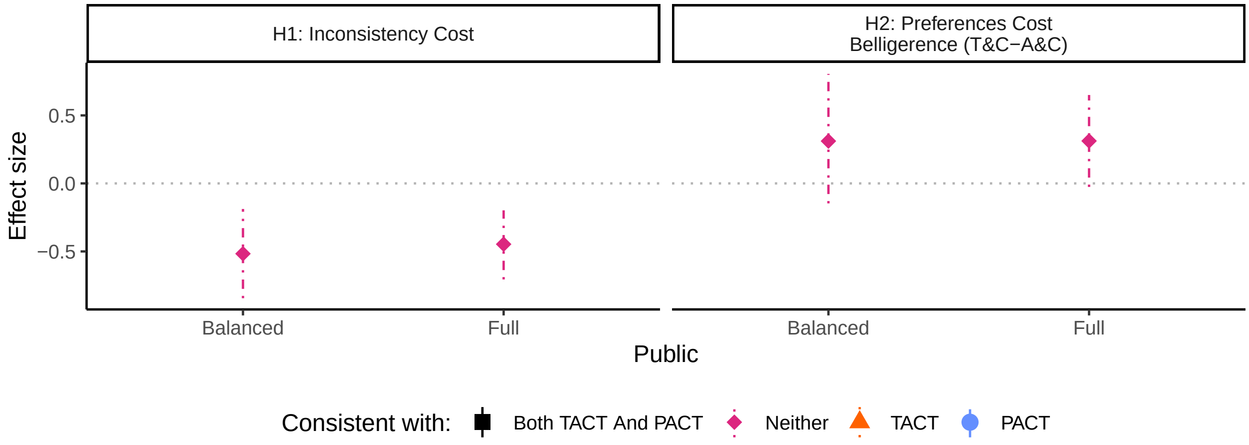

Figure 3: TACT is supported by the mixed publics: inconsistency is punished but preferences are not

[16] A further discussion for each inconsistency cost supports this finding and is available in the appendix.

5.2 New Dynamics in Preference-based Publics

I turn now to the preference-based publics.[17] As hypothesized, I find that hawks and doves enforce different costs on leaders and these are driven by their preferences – not inconsistency. These results are presented in Figure 4.[18]

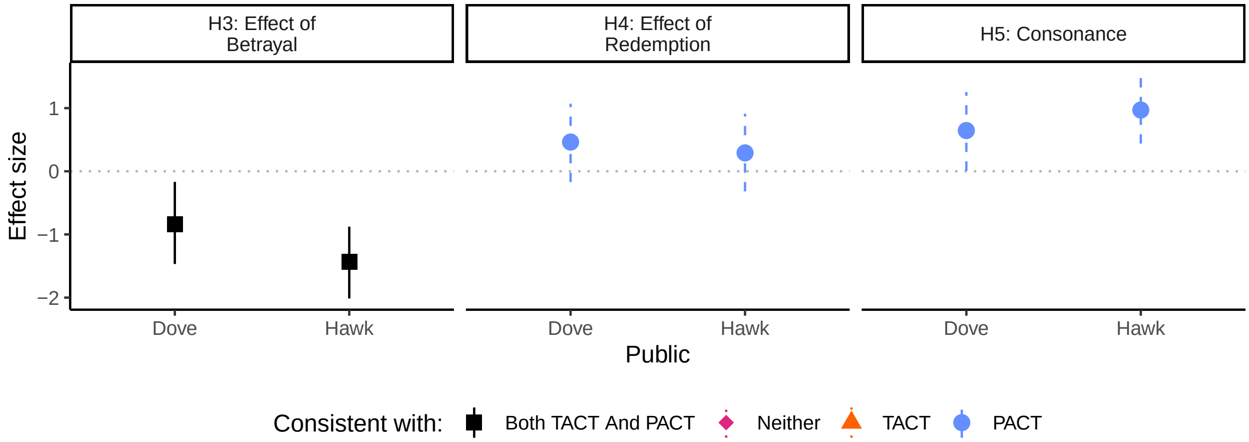

Figure 4: The homogeneous publics behave as PACT predicts: the only inconsistency punished is betrayal and consonance is rewarded

A first difference between the full sample and the preference-based publics is that neither hawks nor doves enforce audience costs on leaders who Back Out (`T&I−A&C`) of a crisis. Though on average backing down has a negative impact on presidential approval ratings for both groups, there is no evidence this effect is significantly different from 0 (hawks: -0.463, $p < 0.072$; doves: -0.197, $p < 0.291$). Additionally, only hawks enforce costs significantly different from 0 on leaders who Back In (`A&I−T&C`). Indeed, leaders who declare they will stay out of a conflict but eventually order intervention have approval ratings 0.68 points lower than those who intervene from the beginning ($p < 0.009$). I cannot find any statistically significant evidence of a reaction from doves when the scenario is reversed ($p < 0.266$), although the effect has the expected sign (-0.197).

Decomposing audience costs into inconsistency and preferences costs further supports PACT. To start, the effects of Betrayal and Redemption are clearly different from one another for both hawks and doves. Betrayal is severely punished; Redemption is not.

Betrayal (H_3): For hawks, approval for the president is 1.434 points lower ($p < 0.001$) when she threatens force but fails to follow through than when she successfully delivers on her threat (`T&I−T&C`). Similarly, for doves, when the president has declared the army will not intervene but eventually orders the troops to engage anyway, her approval ratings

[17] Like before, I find no evidence of correlation between pretreatment variables and treatment assignment.

[18] Another group could theoretically be investigated: the middle group which includes those who are neither hawks nor doves. Doing so, however, presents a conceptual challenge: should this group be expected to include middle-of-the-road indivduals who have a pragmatic approach to the use of military force? Or is this group a mixed constituency that includes both hawks and doves? Because this project is ill-equipped to weigh in on this question, I restrain the scope of this analysis to focus only on hawks and doves.

are 0.842 points lower (p < 0.007) than if she had stayed out of the conflict altogether (`A&I`−`A&C`).

Redemption (H_4): There is no sign hawks punish inconsistency if it yields intervention: the treatment effect of announcing engagement after saying the U.S. will stay out of the conflict compared to staying out of the conflict entirely (`A&I`−`A&C`), is not statistically significant (p < 0.174). The same is true for doves as well: the effect of declaring the army will intervene and not doing so instead of following through (`T&I`−`T&C`) does not appear to be statistically different from zero (p < 0.077).[19]

Consonance (H_5): The effects of the preferences treatments similarly differ and consonance is rewarded. Hawks reward belligerent leaders who intervene compared to leaders who stay out of conflicts (`T&C`−`A&C`; 0.97; p < 0) and doves reward pacifist leaders compared to belligerent ones (`A&C`−`T&C`; 0.645; p < 0.025). By symmetry, this also indicates that the effect of Dissonance is significantly inferior than 0 for both hawks and doves.

Thus, focusing on the hawkish and dovish samples provides extensive support for PACT. As predicted, Betrayal is severely punished (H_3); Redemption does not appear to have a negative impact on presidential ratings (H_4); and the effect of Consonance is significantly greater than 0 (H_5). These dynamics exhibited by the hawkish and dovish samples are consistent with the expectations of PACT. In contrast, these dynamics were not predicted by TACT which expects the effect of Redemption – a type of inconsistency cost – to be negative and of Consonance to be substantively insignificant.

5.3 The Electoral Mechanism and the Vocal Opposition Assumption

The previous section has found strong support for PACT using the approval measure. Now, I show that a direct implementation of the electoral mechanism also supports PACT and implies the need to reconsider the vocal opposition position. Specifically, I find evidence that, because preferences matter, a vocal opposition can result in the leader being rewarded for her inconsistency; and suggestive evidence that a vocal opposition might not actually be required to punish inconsistent leaders at all.

I turn to the two electoral tests outlined previously, focusing first on the challenger who critiques the president's military foreign policy; and then on the challenger who attacks her economic foreign policy.

The results for the foreign policy challenger are presented for the mixed and the preference-based publics in Figure 5. Since this is a direct implementation of the vocal opposition position, it should reproduce the results of the previous sections.

Although the mixed publics reproduce the results found previously for the effect of preferences, I find no evidence in either samples that a leader who is inconsistent is punished. Clearly, these results are at odds with the predictions of both PACT and TACT. Yet, they

[19] Achieving higher power by merging the hawkish and dovish samples allows the identification of a statistically significant but substantively insignificant positive Redemption effect ($Cohen'sd < 0.2$, $p < 0.041$).

could still be consistent with PACT if the cost of Betrayal enforced by one group were to be offset by a positive "cost" (de facto a reward) for Redemption enforced by the other.

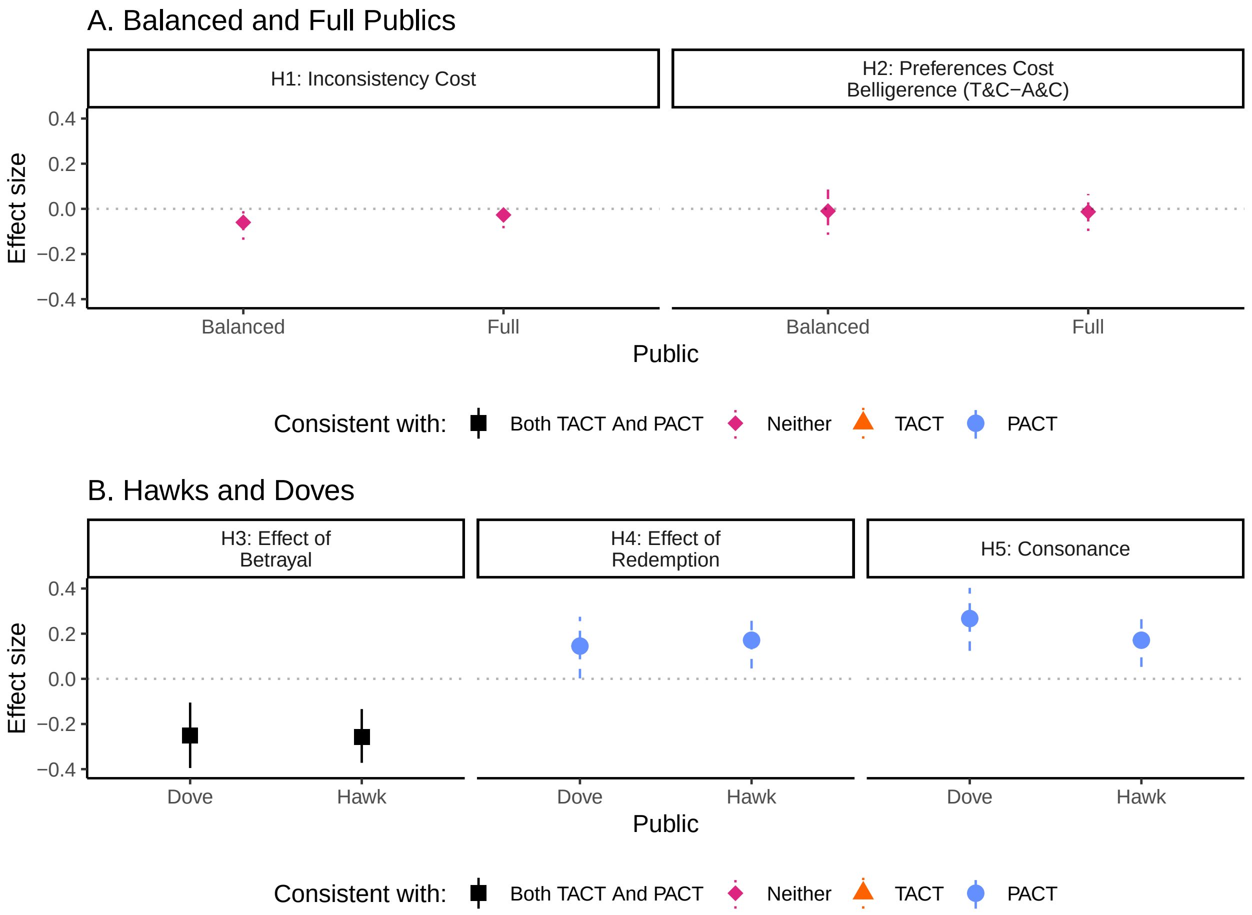

Figure 5: The Foreign Policy Challenger DV supports PACT

The hawkish and dovish samples replicate most of the prior findings. These effects are statistically significant and substantively meaningful for the effect of Betrayal (H_3) and Consonance (H_4).

Unlike before, the effect of Redemption is not only non-negative (H_4) but it is also significantly greater than 0. Hawks reward inconsistency if it yields intervention – if the president orders the army to intervene after saying the U.S. will stay out of the conflict compared to staying out of the conflict entirely (A&I−A&C). Then, the president is rewarded by an increase in vote shares of 17.1% ($p < 0.002$). In the mirror scenario (T&I−T&C), doves reward their president by 14.5% ($p < 0.021$). Pivotal to this result is the fact that any criticism made by the challenger is inherently tied to advocacy for a substantive policy. Thus, it is not surprising a dove whose president backed down from a threat would be willing to reelect this president (despite her initial belligerence) when the alternative is a pro-intervention candidate (and vice-versa for hawks). An important implication is that the presence of an opposition willing to chastise the president for her inconsistency will not always increase the costs she is facing. Rather, the opposition can at times bolster the president's electoral position.

I now turn to the scenario where the president is facing a challenger who criticizes her economic foreign policy but remains silent on the issue of security. The vocal opposition position would expect no particular effect as there is no opposition to exploit the president's inconsistency. This is a difficult test for both PACT and TACT. Because the existence of an opposition willing to criticize the president for her inconsistency has long been presented as a prerequisite for the existence of audience costs, TACT would expect voters should be equally likely to vote for either candidate. Similarly, PACT would expect voters' preferences on economic policies to drive the vote – not those on military foreign policy. Since no information is provided regarding the content of either candidate's economic platform, voters are expected to select the candidate at random. Figure 6A shows that only the full sample is consistent with this view. In fact, the balanced public imposes small but statistically significant costs on inconsistent leaders.

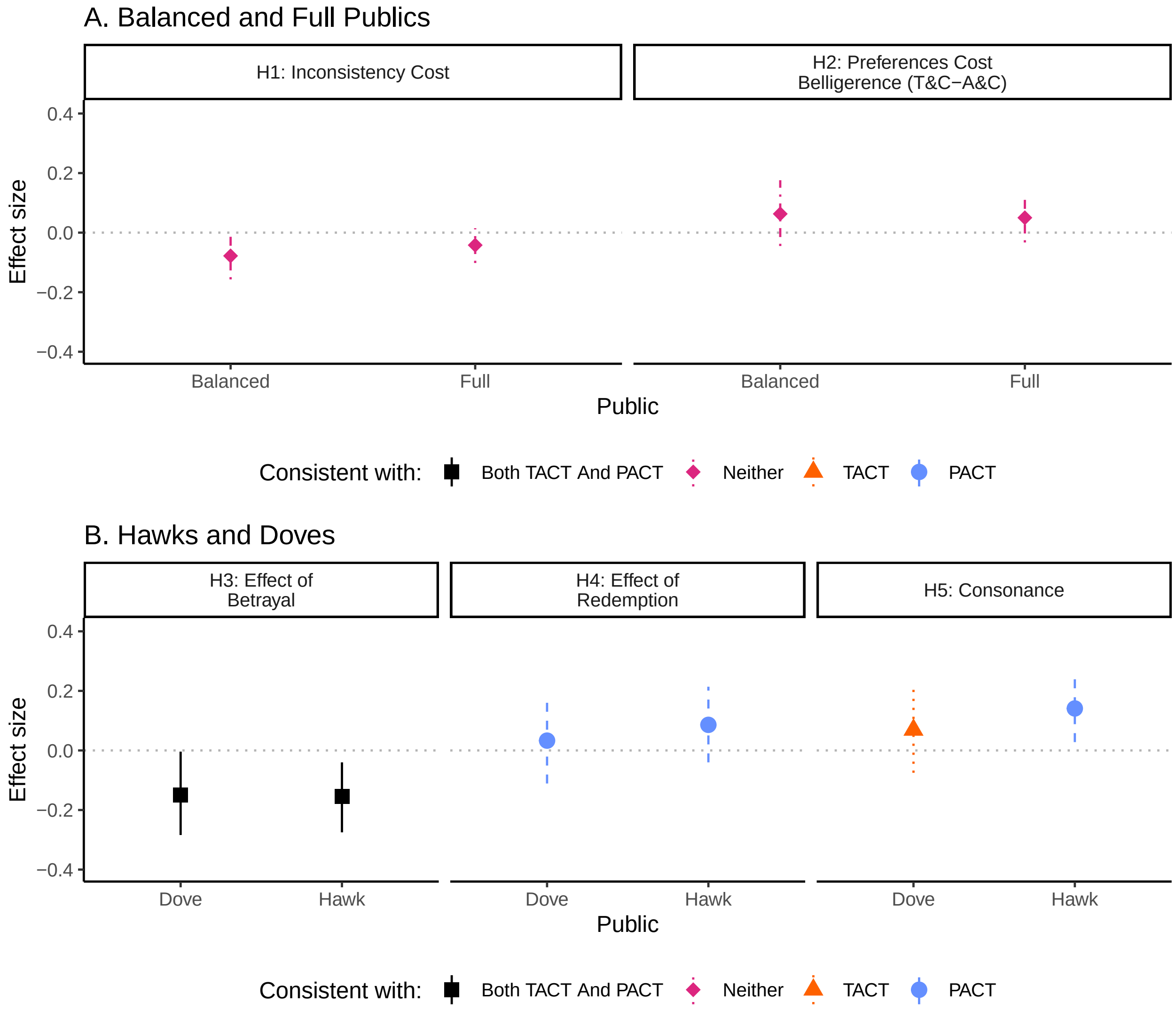

Figure 6: The Economic Challenger DV supports PACT

The situation is also very different for hawks and doves, as shown in 6B. As in the case of the approval and foreign policy challenger measures, Betrayal is severely punished (H_3). Also consistent with PACT, the effect of Redemption is not negative for either group (H_4).

As predicted by PACT, hawks reward their president's Consonance (H_5). For doves, although the effect of Consonance is positive (`A&C`−`T&C`), I find no evidence of it being significantly different from 0 ($p < 0.178$). Although investigating the causes of this divergence would require further research, a possible explanation could be that doves care more about the economy than hawks and thus would conform more closely to the expectation of random voting.

The economic challenger scenario tells us two things. First, Betrayal is costly, always. Even in the absence of a vocal opposition to the policy, individuals will still punish leaders who betray them by voting for the challenger. Second, some individuals are willing to reward and punish leaders based on their preferences, even if they are not offered a viable alternative.

Taken together, these electoral measures support PACT. This is clearly the case when the president is challenged by a foreign policy challenger as all PACT hypotheses are confirmed empirically. Despite all expectations to the contrary, the economic challenger measure also provides some support to PACT. These results suggest that vocal opponents are neither necessary nor sufficient to ensure the punishment of inconsistent leaders.

6 Executive Spin Co-Exists with PACT

Thus far, I have shown that different publics impose different costs on leaders. This discussion would be meaningless, however, if executive leaders could simply bypass these costs by claiming they have received new information. In this section, I demonstrate such executive spin cannot help leaders avoid the costs derived from the audience. Instead, inconsistent and consistent presidents alike benefit from stating they have received new information.

A complementary survey experiment was fielded in the summer of 2020 on a sample of 1172 US citizens of voting age recruited using the MTurk platform. Beyond the executive spin critique, this survey also serves as robustness check for the results discussed previously by way of replication.

Levendusky and Horowitz (2012) argue that the president can shirk punishment after Backing Out by simply announcing they have received new information regarding the crisis. In other words, new information reduces the audience component of TACT to 0. In their response, Levy et al. (2015) are left with puzzling findings: presidential justifications cause the public to reward a president who backs out, but do not appear to change the approval ratings of a president who backs in. Their first cut at explaining this feature is that the public is worried about the impact inconsistency could have on the bargaining credibility of the U.S. but that the evaluations of the president's competence and general concerns about U.S. reputation are skewed. For them, this is a matter of framing between positive and negative commitments.

These findings contradict PACT. If the public cares about the substance of the policy, why would claiming to have received new information matter? Addressing this challenge amounts to retesting hypotheses H_{1-5}.

The experimental design is identical to the one described above with one difference.[20] After learning the attacking country continued to invade its neighbor but before learning the decision made by the U.S. president, participants were told:

> The U.S. president then received new intelligence suggesting involvement [`is / is not`] in America's interests. Military experts agreed that the U.S. [`should / should not`] become involved in this crisis.[21]

A major departure from prior studies is that this excerpt is present for all treatment groups, whereas in prior studies, it was only assigned in combination with inconsistency. Partial assignment is problematic in that it is double-barreled: the consistent president decides on a course of action alone but the inconsistent president makes her decision after receiving both new information and the endorsement of military experts. Since either of these components is expected to favorably impact approval ratings, assigning this treatment to all groups allows us to control for these.

Levendusky and Horowitz (2012) and Levy et al. (2015) find the cost for Backing Out disappears in the presence of new information. Once information is held constant, however, this result disappears. In the full sample, the president is always punished for Backing Out (`T&I−A&C`), regardless of the measure used.[22] Figure 7 depicts audience costs for both Backing In and Out of a crisis. Like Levy et al. (2015), I find no evidence the public would punish Backing In (`A&I−T&C`) in the full sample.[23] The balanced sample complicates this picture: when using the situational approval DV, presidents are punished for both Backing Out and Backing In.[24] However, depending on the challenger, the balanced sample punishes either Backing Out (in the case of the foreign policy challenger) or Backing In (in the case of the economic challenger).[25]

These preliminary results force us to reconsider the executive spin critique: in many cases, new information appears unable to cancel Backing Out and Backing In costs. The variation between these costs, however, remains unaccounted for. I turn once more to the disaggregated data.

The results of this experiment support PACT, as shown in Figure 8. This is unambiguous

[20] Again, I find no evidence treatments were assigned non-randomly, as visible in the appendix.

[21] The president always follows the recommended policy.

[22] Backing Out in the full sample. Approval: -0.355, $p < 0.004$; foreign policy challenger: -0.102, $p < 0.001$; economic challenger: -0.099, $p < 0.004$.

[23] Backing In in the full sample. Approval: $p < 0.188$; foreign policy challenger: $p < 0.326$; and economic challenger: $p < 0.458$.

[24] Backing In $p < 0.006$. Backing Out is significant when using a one-tailed test ($p < 0.048$) and at the margin of the traditional statistical significance threshold using a two-tailed test (-0.884, 0.047).

[25] Foreign policy challenger: Backing Out $p < 0.028$; Backing In $p < 0.41$. Economic challenger: Backing Out $p < 0.272$; Backing In $p < 0.016$.

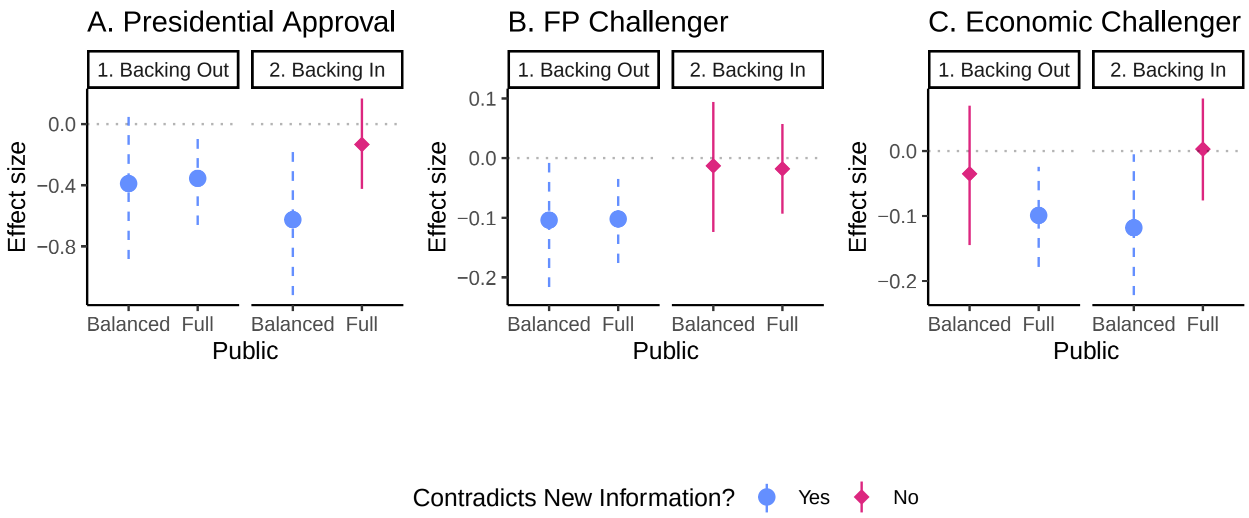

Figure 7: In many cases, Backing Out and Backing In costs remain despite new information

in the case of doves who always punish Betrayal (H_3, A&I−A&C), do not punish Redemption (H_4, T&I−T&C) but reward it, and reward Consonance (H_{3P}, A&C−T&C).

Hawks similarly support PACT. They always punish Betrayal[26] (H_3, T&I−T&C), never punish Redemption (H_4, A&I−A&C), and also reward Consonance (H_5, T&C−A&C) when it comes to approval. There is a caveat, however, as I find no evidence Consonance is rewarded in the electoral measures. This is representative of a broader pattern in the hawkish public: compared to the previous experiment, all effects appear muted.[27] Although further research is needed to ascertain the reason for this difference, a likely cause is the military endorsement included in the treatment.[28] Hawks can be expected to be more sensitive to this endorsement than doves as support for military interventionism is associated with respect for authority at the individual level (Kertzer et al. 2014). Accordingly, it is plausible that hawks are willing to defer more extensively (though not entirely) to the opinion of military experts.

In sum, this experiment replicates the substantive findings of the main experiment and provides evidence in support of PACT. Announcing new information does not reduce the audience component of PACT to 0 but adds a new variable, say k, to the payoffs of the leader. Presidents cannot avoid punishment by claiming to have received new information but rather enjoy a flat competency bonus for doing so. In other words, individuals appear to like presidents who are competent and endorsed by military elites more than presidents who are not. These results refine the findings of Levendusky and Horowitz (2012) and Levy et al. (2015) by showing that this competency bonus co-exists with, rather

[26] The effect of Betrayal with regard to the economic challenger is significantly different from 0 when using a one-tailed test ($p < 0.048$) but not a two-tailed test (-0.228, 0.018).

[27] Although the MTurk hawish public scores lower in militant assertiveness than the Lucid one, this difference is negligeable (−4%).

[28] Levy et al., 2015, also voice this concern. Nonetheless, they run the two-component treatment for the sake of comparability with Levendusky and Horowitz (2012). I do the same for this reason as well.

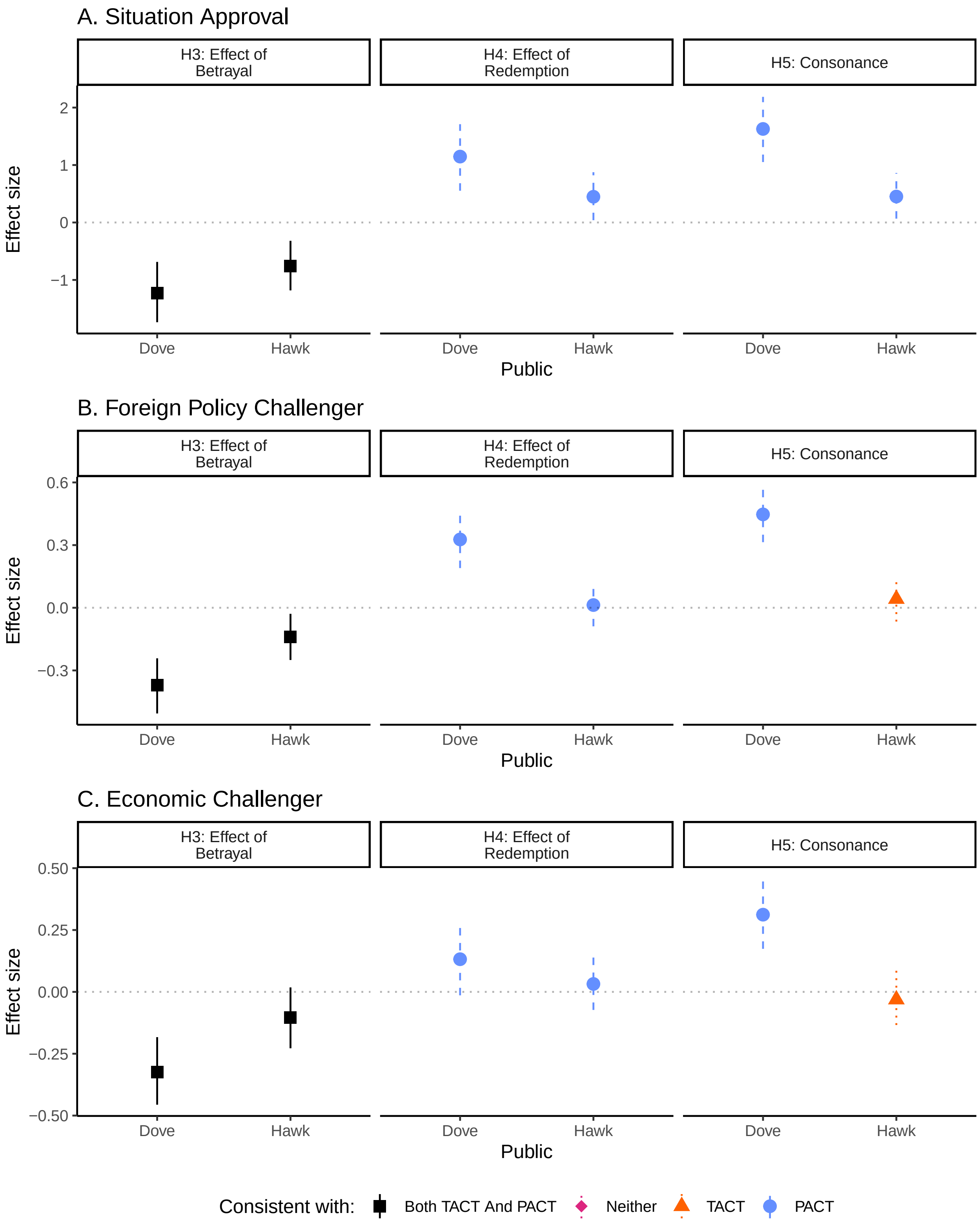

Figure 8: The Executive *cannot* avoid punishment by claiming she has received new information

than undermines, audience costs and PACT.

7 Discussion

This paper critiques how IR thinks about the origins of public opinion and advances our understanding of how individual preferences can affect the strategic incentives of leaders. Its key claim is that, although practical, the assumption of the public as unitary actor has obscured important dynamics tying public opinion to the foreign policy decision-making process, for both heterogeneous and homogeneous publics.

It demonstrated the value of this approach in the context of the audience costs literature by using a game theoretic model that shows preferences, not the dislike of inconsistency, drive the costs publics enforce on leaders. Experimental results provide strong and consistent support for this model. These results reunite audience costs with their critiques and suggest new dynamics of the relationship between public opinion and foreign policy. I briefly discuss three of these here.

First, the substantive choice of policy matters most. Leaders who follow the preferences of the public are rewarded while those who do not are punished. This is consistent with a growing literature that emphasizes the importance of preferences at the expense of inconsistency. Chaudoin (2014) finds individuals with strong preferences over the substance of the policy do not care about inconsistency – only those with weak preferences do. Similarly, McDonald, Croco, and Turitto (2019) conclude policy preferences supersede any consideration of inconsistency – flip-floping – even when policy reversals are made explicit. In contrast with these works, I find that what matters is the interaction between the preferences of the public and the kind of inconsistency: Betrayal is always punished; Redemption is not. This dynamic helps account for the Suez crisis when the British and French backed down from their threat against Egypt yet neither government was punished by their domestic audience. As Snyder and Borghard (2011) argue, this crisis contradicts TACT as inconsistency was not followed by a punishment. However, this is consistent with the intuition of PACT as neither the British nor the French publics supported the conflict (they had dovish preferences) and did not punish Redemption.

Second, the strategic context a leader faces is not static. As the composition of the audience shifts over time, so do the constraints leaders are subject to. These change need not be symmetric in magnitude either. The research presented here suggests that the size of the costs enforced by hawks and doves may not be identical. Uncovering the causes of this variation, whether it is the result of deeper moral values and attitudes or external influences, may be an important key to understanding the strategic context leaders face.

Third, the foreign challenger plays an important role in shaping the audience costs the president must face. Challengers who persevere – regardless of whether the leader emitted a threat or not – force the president to reveal her lie and endure audience costs. In contrast, challengers who would concede when targeted by a credible threat allow the presidents of dovish constituencies to hide their inconsistency in some cases while forcing them to emit threats in others.

Overall this research suggests that the scholarship's skepticism towards the substantive implications of the audience costs literature might not be warranted. More broadly, we may also want to reconsider some of our prior arguments and decide whether they are worth revisiting without abstracting away the nuances of public preferences.

8 References

Barnhart, Joslyn N. et al. 2020. "The suffragist peace." *International Organization* 74(4): 633670.

Baum, Matthew A. 2004. "Going private - public opinion, presidential rhetoric, and the domestic politics of audience costs in us foreign policy crises." *Journal of Conflict Resolution* 48(5): 603–631.

Baum, Matthew A., and Phillip B.K. Potter. 2015. *War and democratic constraint: How the public influences foreign policy*. Princeton University Press.

Broz, J. Lawrence. 2002. "Political system transparency and monetary commitment regimes." *International Organization* 56(4): 861–+.

Bueno De Mesquita, Bruno et al. 2005. *The logic of political survival*. MIT Press.

Busch, Marc L. 2000. "Democracy, consultation, and the paneling of disputes under gatt." *Journal of Conflict Resolution* 44(4): 425–446.

Chaudoin, Stephen. 2014. "Promises or policies? An experimental analysis of international agreements and audience reactions." *International Organization* 68(1): 235–256.

Croco, Sarah E., Michael J. Hanmer, and Jared A. McDonald. 2020. "At what cost? Reexamining audience costs in realistic settings." *The Journal of Politics*: 000–000.

Davies, Graeme A. M., and Robert Johns. 2013. "Audience costs among the british public: The impact of escalation, crisis type, and prime ministerial rhetoric." *International Studies Quarterly* 57(4): 725–737.

Downes, Alexander B., and Todd S. Sechser. 2012. "The illusion of democratic credibility." *International Organization* 66(3): 457–489.

Eichenberg, Richard C. 2005. "Victory has many friends - us public opinion and the use of military force, 1981-2005." *International Security* 30(1): 140–+.

Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American Political Science Review* 88(3): 577–592.

Fearon, James D. 1995. "Rationalist explanations for war." *International Organization* 49(3): Amer Polit Sci Assoc.

Gelpi, Christopher, Peter Feaver, and Jason Reifler. 2005. "Success matters - casualty sensitivity and the war in iraq." *International Security* 30(3): 7–+.

Guisinger, Alexandra, and Alastair Smith. 2002. "Honest threats - the interaction of reputation and political institutions in international crises." *Journal of Conflict Resolution* 46(2): 175–200.

Hainmueller, Jens, and Michael J. Hiscox. 2006. "Learning to love globalization: Education and individual attitudes toward international trade." *International Organization* 60(2): 469–498.

Heffington, Colton. 2016. “Do Hawks and Doves Deliver? The Words and Deeds of Foreign Policy in Democracies.” *Foreign Policy Analysis* 14(1): 64–85.

Herrmann, Richard K., Phillip E. Tetlock, and Penny S. Visser. 1999. “Mass public decisions to go to war: A cognitive-interactionist framework.” *American Political Science Review* 93(3): 553–573.

Holsti, Ole R. 1979. “3-headed eagle - united-states and system change.” *International Studies Quarterly* 23(3): 339–359.

Holsti, Ole R. 2004. *Public opinion and american foreign policy, revised edition*. University of Michigan Press.

Holsti, Ole R., and James N. Rosenau. 1993. “The structure of foreign policy beliefs among american opinion leaders-after the cold war.” *Millennium: Journal of International Studies* 22(2): 235–278.

Hurwitz, Jon, and Mark Peffley. 1987. “How are foreign-policy attitudes structured - a hierarchical model.” *American Political Science Review* 81(4): 1099–1120.

Kertzer, Joshua D., and Ryan Brutger. 2016. “Decomposing audience costs: Bringing the audience back into audience cost theory.” *American Journal of Political Science* 60(1): 234–249.

Kertzer, Joshua D. et al. 2014. “Moral support: How moral values shape foreign policy attitudes.” *Journal of Politics* 76(3): 825–840.

Kreps, Sarah E., Elizabeth N. Saunders, and Kenneth A. Schultz. 2018. “The ratification premium: Hawks, doves, and arms control.” *World Politics* 70(4): 479–514.

Kurizaki, Shuhei, and Taehee Whang. 2015. “Detecting audience costs in international disputes.” *International Organization* 69(4): 949–980.

Leeds, Brett Ashley. 1999. “Domestic political institutions, credible commitments, and international cooperation.” *American Journal of Political Science* 43(4): 979–1002.

Levendusky, Matthew. 2009. *The partisan sort: How liberals became democrats and conservatives became republicans*. University of Chicago Press.

Levendusky, Matthew S., and Michael C. Horowitz. 2012. “When backing down is the right decision: Partisanship, new information, and audience costs.” *Journal of Politics* 74(2): 323–338.

Levy, Jack S. et al. 2015. “Backing out or backing in? Commitment and consistency in audience costs theory.” *American Journal of Political Science* 59(4): 988–1001.

Mattes, Michaela, and Jessica L. P. Weeks. 2019. “Hawks, doves, and peace: An experimental approach.” *American Journal of Political Science* 63(1): 53–66.

McDonald, Jared, Sarah E. Croco, and Candace Turitto. 2019. “Teflon don or politics as usual? An examination of foreign policy flip-flops in the age of trump.” *Journal of Politics* 81(2): 757–766.

Mercer, Jonathan. 2012. “Audience costs are toys.” *Security Studies* 21(3): 398–404.

Mueller, John E. 1973. *War, presidents, and public opinion*. New York: John Wiley.

Nomikos, William G, and Nicholas Sambanis. 2019. “What is the mechanism underlying audience costs? Incompetence, belligerence, and inconsistency.” *Journal of Peace Research* 56(4): 575–588.

Quek, Kai. 2017. “Type ii audience costs.” *Journal of Politics* 79(4): 1438–1443.

Rathbun, Brian C. et al. 2016. “Taking foreign policy personally: Personal values and foreign policy attitudes.” *International Studies Quarterly* 60(1): 124–137.

Rosenau, James N. 1961. *Public opinion and foreign policy : An operational formulation.* Random House, Inc.

Schultz, Kenneth A. 2001. “Looking for audience costs.” *Journal of Conflict Resolution* 45(1): 32–60.

Schultz, Kenneth A. 2005. “The politics of risking peace: Do hawks or doves deliver the olive branch?” *International Organization* 59(1): Int Studies Assoc.

Slantchev, Branislav L. 2006. “Politicians, the media, and domestic audience costs.” *International Studies Quarterly* 50(2): 445–477.

Smith, Alastair. 1998. “International crises and domestic politics.” *American Political Science Review* 92(3): Peace Sci Soc.

Snyder, Jack, and Erica D. Borghard. 2011. “The cost of empty threats: A penny, not a pound.” *American Political Science Review* 105(3): 437–456.

Tomz, Michael. 2007. “Domestic audience costs in international relations: An experimental approach.” *International Organization* 61(4): 821–840.

Trachtenberg, Marc. 2012. “Audience costs: An historical analysis.” *Security Studies* 21(1): 3–42.

Trager, Robert F., and Lynn Vavreck. 2011. “The political costs of crisis bargaining: Presidential rhetoric and the role of party.” *American Journal of Political Science* 55(3): 526–545.

Weeks, Jessica L. 2008. “Autocratic audience costs: Regime type and signaling resolve.” *International Organization* 62(1): 35–64.

Wittkopf, Eugene R. 1990. *Faces of internationalism: Public opinion and american foreign policy*. Duke University Press.