The Puzzle of Coercion Failure: How Psychology Explains Resistance to Threats

Kathleen E. Powers,*, Dan Altman.†

August 12, 2019‡

Draft manuscript prepared for presentation at the 2019 Annual Meeting of the American Political Science Association. We welcome comments, but kindly don't circulate.

Abstract: Leaders often turn to coercive diplomacy to get what they want, yet compellent threats have a surprisingly poor success rate. Targets stand firm, rather than back down, even in the face of powerful challengers. Drawing from research in psychology, we identify one important yet unrecognized factor that causes people to resist threats: psychological reactance. Reactance theory explains that when someone perceives a personal threat to their freedom to choose, they will respond by trying to restore their autonomy. They will perform a forbidden action rather than capitulate. We test our theory with a novel experiment designed to isolate the effect of reactance by comparing how people respond to personal threats or costly but impersonal constraints. Controlling for prominent alternative explanations, we find evidence that people who are the target of a direct threat are less likely to capitulate and more likely to promote aggression against their opponent — and that this effect is primarily mediated by a person's belief that they are being manipulated. Coercion failure has a psychological foundation.

*Assistant Professor, Department of Government, Dartmouth College. kathleen.e.powers@dartmouth.edu. http://kepowers.com

†Assistant Professor, Department of Political Science, Georgia State University. daltman@gsu.edu. http://www.danielwaltman.com/

‡We thank Danielle Lupton, Jonathan Renshon, and workshop participants at Dartmouth College and Georgia State University for helpful comments.

1 Introduction

Why do actors often stand firm in response to coercive threats? In February 1991, Iraq had invaded and occupied Kuwait, leading to a U.S.-led coalition's aerial assault designed to drive them out. Unsatisfied by a diplomatic stalemate, President Bush issued an ultimatum: "The coalition will give Saddam Hussein until noon Saturday to do what he must do—begin his immediate and unconditional withdrawal from Kuwait." As the Chicago Tribune summarized, the coalition's directive was to "Get out, or else."[1] Saddam did not capitulate, choosing to continue the war in spite of long odds and heavy costs (Herrmann, 1994), nor is his intransigence unusual — leaders defy coercive threats so often that Art and Greenhill (2018, 78) conclude that the "track record" for coercion is at best "underwhelming."

International relations scholars have proposed a panoply of theories to explain both the puzzlingly high failure rate of many different types of coercion across a diverse array of cases and eras. For instance, perhaps Saddam would have given in to the proposed terms if he knew that the U.S. had a superior battle plan, had correctly perceived U.S. interests and Bush's resolve to stamp out aggression (Duelfer and Dyson, 2011), or if he could have been assured that he would not be stripped of power and punished by domestic audiences upon backing out (Fearon, 1994). While it is possible that leaders make cold calculations about credibility and costs when confronted with coercion (Sechser, 2010), a multitude of separate rationalist models are ultimately unsatisfying as explanations for such a widespread phenomenon. If resistance and retaliation are the norms rather than exceptional responses (Sechser, 2011; Altman, 2017), our theories should treat them as such. Whereas rationalist accounts are optimistic that compellent threats can succeed when the situation is right, "[t]he psychology of coercion is challenging, subtle, and often futile" (McDermott, 2017, 91).

We draw from psychological research to propose a reactance-based theory of coercion failure (Brehm, 1966). Imagine that someone tells you what to do, and threatens you with costs if you do not abide. Do you dispassionately weight the costs and benefits of acquiescence? Or do you react to the diktat by digging in? Perhaps you favored a certain university before your parents stated that you must attend or forego their financial assistance, when it became less attractive. Or sensing that you have been manipulated, you become angry and seek punitive retaliation. A leader might alienate

[1] George H.W. Bush qtd. in George de Lama and Timothy J. McNulty, 1991. "U.S. Ultimatum to Iraq: Get Out or Else," *Chicago Tribune*, 23 February.

an ally for refusing to entertain his proposal to purchase territory, for example.[2] Psychological reactance theory tells to expect these responses — when a person believes that another actor is trying to constrain her autonomy, she will try to restore her freedom through resistance, defiance, and aggression (Miron and Brehm, 2006).[3] In short, we argue that leaders resist coercion — and sometimes retaliate (Dafoe, Hatz and Zhang, 2017) — because threats promote obstinance.

We present a novel experimental design that manipulates the presence or absence of a coercive threat but holds constant other key factors — including the direct costs of defiance. Our first study, which presents survey respondents with a hypothetical naval standoff where they must choose whether to risk their soldier's lives to reclaim a valuable piece of territory or to turn back, reveals three important results. We find that people exposed to a coercive threat are less likely to turn back and more willing to retaliate against their challenger, compared to a group of people who face an impersonal risk from a storm. People react to the threat, not just to prospective costs and benefits of action. Second, we find that coercive threats have direct effects on key mechanisms associated with psychological reactance — anger and negative cognitions — but not on factors associated with rationalist explanations. People do not update their beliefs about the opponent's relative power, for example, in response to a threat. Third, we leverage causal mediation analysis to show that while direct threats can also arouse reputation concerns, the effect of coercion on resistance primarily runs through a person's belief that she is being manipulated. We close with a proposal for a follow-up study that manipulates coercion in a different policy context.

2 The Puzzle of Coercion Failure

The balance of research suggests that coercive threats fail to achieve desired outcomes more often than they succeed, across categories from economic sanctions to conventional military and nuclear threats (Art and Greenhill, 2018). This does not mean that coercion never succeeds: In 1938, Czechoslovakia relinquished the Sudetenland in compliance with Hitlers ultimatum at Munich. In

[2] Maggie Haberman, 2019. "Trump says Danish Prime Minister was 'Nasty' in Rejecting Greenland Offer," *The New York Times*, 21 August.

[3] Of course, ours is not the first attempt to integrate psychology with resistance to compellence. For example, some scholars focus on cognitive phenomena like loss aversion (Berejikian, 2002*a*; Jervis, 1992) to explain why coercion (compellence) is harder than deterrence, the roles played by anger and pride (Hall, 2011, 2017; Marwicka, 2018), or a general logic of reputation and honor (Dafoe, Hatz and Zhang, 2017). We complement this research by proposition a theory that bridges the illusory divide between "cold" and "hot" cognition (see Kertzer and Tingley, 2018 for a brief review of hot and cold cognition in IR).

1962, U.S. threats led the Soviet Union to withdraw nuclear missiles from Cuba.[4] Armed groups drove the United States out of Lebanon in 1984. Economic sanctions contributed to Irans nuclear concessions in 2015. Yet these landmark cases of undisputed coercive success are remarkable for their rarity. Measured against the broader history of coercive diplomacy, failure is the norm.

We use coercion to describe an explicit demand by one actor (the challenger) that another actor (the target) alter the status quo in some material way, backed by a threat to impose costs if the target does not comply (Sechser, 2011, 379).[5] This definition places threat-making at the heart of coercion. Following some - but not all - of the literature, we use the term coercion to refer to what Schelling (1966) labeled compellence.Whereas deterrence seeks to preserve the status quo, compellent threats demand changes to it.[6] It also incorporates a punishment into the definition; this sets aside coercion by denial, which entails a threat to undermine and defeat the adversarys strategy rather than to inflict harm (Pape, 1996). Coercion failure occurs when the target of a threat does not comply with the demand, irrespective of whether the coercer then carries out the threat.

Coercion failure is not limited to a single form of statecraft, and IR scholars find low rates of capitulation across domains and challenger type. According to the Militarized Compellent Threats dataset, 39 percent of 242 explicit coercive threats from 1918 to 2001 succeeded in eliciting compliance from the target (Sechser, 2011). Studies of specific issue areas suggest a lower rate of success. Willard-Foster (2018) reports that states are rarely able to achieve regime change via coercive threats, leading them to instead resort to covert and overt interventions.[7] Although regime change may be too high a cost, Altman (2017) reports that states have only rarely acquired territory by making threats, and most of those successes are confined to a cluster preceding World War II. Instead, states more often seized territory by fait accompli. Even in territorial disputes over small islands and border regions, threats rarely acquire territory.

Aerial bombardment can inflict catastrophic suffering on states and their populations and was prominent during the Second World War, yet it has rarely (if ever) coerced a regime into altering the status quo. Instead, concessions came after progress on the ground.Leaders may be insulated against the worst effects of an aerial assault, but research shows that civilians do not demand that their government give in. Far from turning against the leadership, civilians "rally around the flag"

[4] Along, of course, with the secret concession to withdraw Jupiter missiles from Turkey.

[5] This definition is taken almost verbatim from Sechsers definition of militarized compellent threats. However, we altered the language to include non-state and state actors alike, and other costs beyond force.

[6] We recommend the question of whether reactance applies as potently to deterrence as an area for future research.

[7] Also see O'Rourke (2018); Downes and O'Rourke (2016).

to support the leader when they are faced with low-level bombardments, or turn their attention to day-to-day survival against high-level assaults (Pape, 1996; Horowitz and Reiter, 2001). Members of the public support resistance or ignore politics when they are most vulnerable — but do not call for their leaders to give in to the threat.

Engaging with many of the same cases, scholars of war termination ask why wars tend to persist even after it becomes clear what the outcome will be. Often, the losing side equivocates before capitulating or refuses to surrender (Goemans, 2000; Reiter, 2009; Weisiger, 2013). That reluctance to concede is an incarnation of the broader puzzle of coercion failure.

The modern proliferation of research on terrorism also shows that threats do little to extract concessions from their targets.[8] Analyzing 648 terrorist groups, Jones and Libicki (2008) find a ten percent success rate and doubt that terrorism caused contributed greatly to causing many of those ostensible successes. Abrahms (2006) identified only two terrorist groups with success levels beyond tactical gains such as prisoner releases. One of these, the Tamil Tigers, was later defeated. Although Pape (2003) reports a high success rate for suicide terrorism – 50 percent – Ashworth et al. (2008) and Krause (2013) challenge the validity of that conclusion. Comparing rebel groups that use terrorism to rebel groups that eschew it, Fortna (2015) reports that the groups using terrorism succeed less often.

The prevalence of coercion failure is not limited to violence — economic statecraft and cyber threats have similarly poor track records. Estimates vary regarding the precise rate at which economic sanctions coerce a policy change. Whereas Pape (1997) estimates a low of 4%, Hufbauer, Schott and Elliott (1990) suggests that sanctions achieve their goals 34% of the time.[9] More recent and comprehensive data on 1,412 cases of economic sanctions and verbal threats to impose them finds a 27 percent success rate (Morgan, Bapat and Kobayashi, 2014; Drezner, 1999).[10] The overall picture remains bleak. Sanction threats are more likely to change policy in a highly contingent set of circumstances (Drezner, 2011). Importantly, those figures describe observed outcomes after sanctions and threats of sanctions, not a causal effect of sanctions (Nooruddin, 2002). Cyber coercion is new, but pioneering research on the topic already suggests that that it seldom generates concessions (Lindsay, 2013; Gartzke, 2013). Finally, Greenhill (2010) reports that coercive engineered migration

[8] Kydd and Walter (2006) argue that terrorist groups have a variety of strategic aims, such that some threats or uses of force are intended to e.g., provoke their target rather than encourage them to back down. Our discussion focuses on cases that are consistent with our definition of coercion, and not intentional provocation or other aims.

[9] In most of the disputed cases, the sanctions themselves failed to garner a concession and the coercer then imposed their will with military force.

[10] This figure rises if negotiated settlements are included or cases with outcomes coded as missing are removed.

succeeds in full 57 percent of the time and succeeds in part 73 percent of the time. We return to this anomaly in later discussion.

2.1 Existing Explanations

Why is it so hard to compel a state to change their behavior? This is not a puzzle driven by powerful actors who refuse to acquiesce to demands from weak challengers — U.S. resistance to a threat from Belize is easy to explain. To the contrary, we know that strong states find it difficult to coerce relatively weak opponents.[11] On one hand, weak actors might offset their lack of power with greater resolve (Mack, 1975). Due to selection effects, crises and disputes tend to break out despite power asymmetry only when the weak have high resolve or another countervailing advantage (Fearon, 2002). On the other hand, Arreguin-Toft (2001) argues that weak actors are more likely to use barbaric strategies that exploit powerful states' unwillingness to respond in kind, and Haun (2015) argues that weak actors are left with little choice when powerful challengers make demands that threaten their survival. Similarly, evidence that nuclear powers are able to coerce non-nuclear states via nuclear blackmail is thin (Betts, 2010; Sechser and Fuhrmann, 2017).[12]

Many of the most compelling and influential theories of coercion failure apply only to one form of coercion. Economic sanctions are thought to fail because the target can substitute a new trading partner for the coercer, rather than losing the trade altogether (Bapat and Morgan, 2009; Early, 2015).[13] Nuclear coercion fails because threats to use nuclear weapons over all but the gravest issues are not credible (Sechser and Fuhrmann, 2017). Terrorism fails because targets view terrorists as implacably hostile and thus doubt that concessions would lead reduce the change of another attack (Abrahms, 2006). The strong may fail to coerce the weak because the balance of resolve offsets the balance of power (Mack, 1975). The literature abounds with such specific explanations that have advanced our understanding of the conditions under which particular coercive threats should succeed or fail.

Although individually compelling, theories of coercion failure that are confined to one of the many subsets of coercion are collectively unsatisfying. By analogy, suppose that forms of coercion

[11] Indeed, the body of research dedicated to understanding why the strong find it so difficult to coerce the weak underscores the broader puzzle of coercion failure.

[12] But see Kroenig (2018).

[13] Early suggests that third-party sanctions busters sometimes undercut sanctions for strategic reasons or even provide offsetting aid to sanctioned states. While modern sanctions strategies can harness the global financial system to limit outside options, analysts are skeptical that they are necessarily better at extracting concessions than "old" comprehensive sanctions (Feaver and Lorber, 2015).

are like engineers. Imagine that ten engineers each attempt to build a bridge based on a specific design; all ten bridges then collapse. One engineer may misunderstood the plans. Another made a mistake at an inopportune moment in the process. But given that all ten failed, it seems to prudent to first ask whether the design itself was faulty. Because most types of coercion fail most of the time, we propose that a powerful psychological reluctance to capitulate better fits the sweeping extent to which targets meet threats with resistance.

To be sure, there is no shortage of general explanations for why threats fail. These include: 1) excessive demands, 2) insufficiently costly punishments, 3) a dearth of credibility, 4) a coercer with too little power relative to the target, 5) loss aversion, and 5) threats that are not accompanied by a credible assurance that compliance will avert punitive action (Schelling, 1966; Slantchev, 2005; Powell, 1990; Jervis, 1992).[14] Leaders might also refuse to submit out of fear of inviting future threats (reputation) or seeing their political fortunes suffer domestically (audience costs) (e.g., Schelling, 1966; Mercer, 2010; Press, 2005; Dafoe, Renshon and Huth, 2014; Fearon, 1994; Snyder and Borghard, 2011).[15] Powerful evidence supports each of these claims, and we do not argue that they are irrelevant factors. Instead, we believe that these explanations are more compelling as reasons why particular threats failed than why threats fail in general. Why cannot coercers demand less, threaten more, or signal resolve more emphatically? Why would it be so difficult for coercers to create a cost-benefit calculation for targets more conducive to threats succeeding? We propose that even if actors could account for each of these variables in calibrating their threat, a psychological constraint against capitulation may nevertheless cause the target to stand firm.

3 A Reactance-Based Theory of Coercion Failure

We argue that coercive threats often fail to extract concessions from their targets – and occasionally backfire (Dafoe, Hatz and Zhang, 2017) – because humans have a psychological aversion to perceived limits on their freedom to act (Brehm, 1966). Psychological reactance theory explains that when an individual's autonomy is threatened — when she feels like she is being manipulated, backed into a corner, or constrained — she will experience the "motivational state of reactance" (Miron and Brehm, 2006, 4). This "unpleasant" state comprises a combination of emotional arousal and

[14] See Art and Greenhill (2018) for a comprehensive recent review.

[15] Indeed, we regard reputation and audience cost incentives as the most compelling rationalist answers to the puzzle of coercion failure and address them in several ways in our research design.

negative cognitions (Gadarian, 2014, 69), including anger and derogation directed at the source of the threat.[16] Insofar as coercive threats entail attempts to get an actor "to do something it does not want to do" (Art and Greenhill, 2018, 78), we expect targets to experience reactance.

When a person feels that her freedom to choose is threatened — when she experiences reactance — she will respond by trying to restore her freedom. She may resist, derogate her manipulator, deny the threat exists, and do the opposite of what she has been told rather than capitulate (Dillard and Shen, 2005, 146; Katz, Byrne and Kent, 2017). In other words, she will perform the "forbidden act" and reaffirm her commitment to a belief or policy rather than acquiesce. Anyone who has interacted with a toddler can confirm that when told they must wear rain boots to play outside, they will defiantly slip into a pair of sandals and run out the door (Miron and Brehm, 2006). Notably, reactance occurs even when the target of a threat is not the actor's most preferred option. The child may love her rain boots the most, but she will rate the sandals as more attractive when they are proscribed.

Reactance theory is prominent in psychology and communications research, but to our knowledge it has not been implicated in IR scholarship on coercion. Brehm's (1966) monograph has been cited more than 7,500 times and reactance "has now become so well-accepted by experimental psychologists" that applications abound (Laurin et al., 2013, 153). Scholars find evidence for reactance effects in contexts ranging from legal punishments (Moore and Pierce, 2016) to cigarette warning labels (Erceg-Hurn and Steed, 2011; LaVoie et al., 2017), alcohol consumption (Dillard and Shen, 2005), censorship (Behrouzian et al., 2016), and beliefs about whether handguns should be banned (Miller et al., 2013). Vrij et al. (2017) propose reactance as one explanation for why "enhanced interrogation" techniques are often ineffective. In political science, reactance informs research on why people resist persuasive communications. It is one of the theoretical explanations for the so-called "boomerang" or "backfire" effect, for example, whereby people respond to attitude inconsistent information by doubling down on their prior beliefs rather than succumb to efforts to manipulate their freedom of thought (Peffley and Hurwitz, 2007; Nyhan and Reifler, 2010; Nisbet, Cooper and Garrett, 2015). Germane to IR, Gadarian (2014) finds evidence that vivid, threatening messages about potential terrorist attacks can provoke reactance among contra-partisans — Democrats who viewed a news story that depicted an urgent terrorist threat were less likely to support the Bush

[16]For example, if the threat can be attributed to a specific person, recipients rate them lower on post-treatment evaluations. In the case of persuasive messages that people believe are designed to tell them what to think or believe, they will list more counter-arguments compared to when they read a less heavy-handed message.

administration's counterterrorism policies than those who viewed a more subdued version of the story.

We focus on psychological reactance for three reasons. First, reactance theory applies when people believe that they have the right to decisional autonomy (Quick and Stephenson, 2007). For states and leaders in particular, who have de jure and de facto sovereignty (Krasner, 1999) — but also for non-state targets like rebel groups who believe they have the freedom to engage in violence, control the population, or build a state — following a challenger's directive means ceding some autonomy. Reactance is less likely if a target perceives the challenger to have legitimate authority over them, like EU member-states submitting to rulings from the European Court of Justice. While challengers might appeal to international norms or institutions, targets do not universally submit to these supranational authorities. Insofar as North Korea believes that they have the sovereign right to build nuclear weapons, statements to the contrary will be met with resistance. The fact that reactance is a response to an absence of authority, and not power, can help explain why a challenger's military superiority is not correlated with success (Sechser, 2011, though see Sechser (2010) for a reputation-based explanation for this anomaly and Fearon (2002) for a selection effects explanation).

Second, reactance is most likely when challengers use forceful, dogmatic language (Quick and Considine, 2008; Gadarian, 2014; Steindl et al., 2015). People respond positively to "nudges" (Thaler and Sunstein, 2009), but not to mandates or statements that they *must* make a particular choice. Because compellent threats, by definition, include "a demand for a material change in the status quo" (Sechser, 2011, 380) they are most likely cases for reactance. The same language that signals credibility, clarity, and commitment amplifies the chance that a challenger will instead harden their target against them. George W. Bush, for example, declared in a public broadcast that "Saddam Hussein and his sons *must* leave Iraq within 48 hours. Their refusal to do so will result in military conflict, commenced at a time of our choosing" (emphasis added).[17] Bush's communication is clear and direct, but like other coercive demands it employs the threat-to-choice language that prompts reactance (Dillard and Shen, 2005).

Third, psychological reactance encompasses a range of behaviors and emotions associated with coercive diplomacy failures and subsumes other psychological explanations in the IR literature. As

[17] George W. Bush, 2003. "President Says Saddam Hussein Must Leave Iraq Within 48 Hours," 17 March. Transcript available at https://georgewbush-whitehouse.archives.gov/news/releases/2003/03/20030317-7.html.

noted above, reactance predicts that individuals might respond to a threat with direct attempts to restore their freedom. This can include standing firm and continuing to perform the forbidden behavior — maintaining a WMD program in spite of resolutions forbidding it — or with direct aggression against the challenger (Miron and Brehm, 2006). It thus explains why coercive threats often not only fail, but provoke violent backlash (Dafoe, Hatz and Zhang, 2017; LaFree, Dugan and Korte, 2009). It also complements research on emotions in IR. Anger (Marwicka, 2018), sometimes caused by personal outrage at violations to a leader's core values (Hall, 2017), encourages actors to stand firm or fight back, and to accept greater risks. Anger is a key mechanism for the link between reactance and defiance. Finally, the reactance response is greatest when it is difficult but not impossible for an actor to restore her lost freedom and can explain why faits acompli often succeed where threats would fail (Miron and Brehm, 2006; Altman, 2017). When a freedom has been lost — when concessions are taken rather than extracted via coercion — reactance is muted.

We expect that the leader of a target state will experience reactance when she receives a direct threat from another actor who does not otherwise have authority to shape her policies. The primary implication of our theory is that even when resisting the threat is costly, a leader will stand firm rather than back down. Because the banned alternative becomes more attractive during reactance, a leader will absorb costs that they otherwise would not in order to reassert their autonomy. A direct threat will also increase the probability that a person chooses not only to resist the threat but to attack back — meeting their challenger with sanctions or a military strike. This leads to our first two hypotheses:

Hypothesis 1: The target of a coercive threat will resist compliance.

Hypothesis 2: The target of a coercive threat will advocate an aggressive response to their challenger.

Second, we draw from research showing that reactance operates through two key emotional and cognitive pathways (Dillard and Shen, 2005) to develop and test hypotheses about the causal mechanisms. Specifically, reactance should increase anger and negative attitudes toward the challenger:

Hypothesis 3: The target of a coercive threat will be angry at the challenger.

Hypothesis 4: The target of a coercive threat will evaluate their challenger negatively.

By identifying and measuring indicators associated with reactance, we can better estimate the

degree to which coercive threats cause resistance due to psychological reactance versus alternative mechanisms. Our argument complements some of the psychological mechanisms through which concerns about reputation and honor operate (Dafoe, Renshon and Huth, 2014), for example, but we estimate alternative explanations alongside reactance indicators in order to evaluate whether ceteris parabis, reactance matters.

4 Research Design

We conducted an online survey experiment in August 2019 to test our theory that reactance provides an important psychological constraint against capitulation under threat. In Study 1, participants take the position of a leader involved in a territorial dispute with a hypothetical opponent and must choose whether to risk their own soldier's lives to retain their land. In Study 2 (proposed follow-up), participants pose as advisers to the U.S. president and must decide whether to continue supporting a pro-democracy movement protesting an oppressive dictator — a choice that comes with a risk of a terrorist attack on U.S. soil — or to cease support for that pro-democracy group. A common structure unites the two studies: Participants face a choice about whether to risk material costs to continue a valuable course of action, but their exposure to a direct threat seeking to dictate that choice to them is randomly assigned.

These original experiments are designed to meet two important challenges to testing our causal claims. First, experiments offer a degree of control that is difficult to match with observational data. Process-tracing methods can implicate psychological factors in a single case (Marwicka, 2018), but the sheer volume of alternative explanations for coercion failure can easily mask the role played by reactance (Art and Greenhill, 2018). Moreover, we are interested in testing a catholic theory about how people respond to threats rather than explaining why Milosevic refused peace at Rambouillet or why the Taliban refused to turn Osama bin Laden over to the U.S. in particular.

Yet even experimental studies can threaten causal inference insofar as researchers fail to account for all reasonable alternatives or when "information leakage" allows participants to impute additional characteristics onto the targeted construct (Dafoe, Zhang and Caughey, 2018). Suppose, for instance, that respondents exhibited surprising reluctance to concede part of a disputed territory under threat of war. Reactance might be responsible for their recalcitrance, but the effect is indistinguishable from other explanations. Perhaps respondents place enough value on the territory to outweigh the

risks, the threat is not credible, or they worry that capitulation now will encourage the adversary to take the remainder of the disputed territory later on. The experimental vignettes must control for all other reasons that a target might resist their challenger.

Second, experiments remain the gold standard for testing causality (Hyde, 2015), but coercive threats are endogenous to their circumstances and lack clear counterfactuals that pose a challenge to controlled comparison (Fearon, 2002). Consider the following example. Suppose that we want to know how people respond when threatened with a coercive bombing campaign akin to NATO's intervention in Serbia. Creating a coercion treatment is straightforward: we can tell one group of participants that another actor threatened to drop bombs on their territory unless they make an important concession. The relevant comparison is more elusive. Exposing one group of participants to a positive inducement rather than the threat, for example, would change how the problem is framed and the costs associated with resistance (Jervis, 1992; Berejikian, 2002*b*). A content-free control would lose all meaning insofar as the threat constitutes the scenario: if we found that people were less likely to make voluntary concessions than to give in to a threat, we could only conclude that costs are relevant to a decision. But without the bombs, the scenario is fundamentally different.

What constitutes the non-coercion equivalent to a bombing threat that is the same in every way except that it does not involve coercion? Research on reactance is instructive in this respect. Psychologists propose that "impersonal" threats inspire less reactance than "personal" threats: participants who expect to choose a prize among five options are more hostile when a researcher deliberately removes one alternative than when they are told that it was lost in the shipment process (Cherulnik and Citrin, 1974; Brehm and Brehm, 2013; Miron and Brehm, 2006). "Nature" could drop bombs on certain types of Serbian targets after e.g., a war game gone awry, a natural disaster, or a failed transport mission,[18] but this strains credulity for most scenarios that entail a coercive threat.

Worse still, a natural disaster with damage comparable to a bombing campaign does not completely deal with the counterfactual challenge. We are interested in whether a target decides to dig in after receiving a threat but before the punishment — if the challenger uses brute force, the target does not have the opportunity to avoid paying costs. For the decision environment to remain the same, respondents in Serbia's position would need to believe that the "natural disaster" version of a bombing campaign would only occur if they continue their policy of ethnic cleansing in Kosovo but

[18]See Dafoe, Zhang and Caughey's (2018) research on "embedded natural experiments" for a complete discussion of the importance of information equivalence in survey experiments.

not otherwise. We do not know of any weather events that are so discerning. The vast majority of coercion scenarios cannot be used to experimentally isolate the effect of reactance.

To overcome these challenges and leverage the internal validity gains offered by experiments, we design coercion vignettes with two atypical characteristics. First, the damage (threatened punishment) could occur either due to deliberate action by the challenger or due to a realistic but impersonal alternative mechanism — what we refer to as *natural costs*. Second, whether or not that damage occurs is determined by a decision on the part of the victim (coercive target) about whether to lose something valuable but avoid paying costs through capitulation.

For example, suppose a state is weakening its neighbor by supporting a rebel group. That neighbor could intentionally expel refugees to the rebel group's sponsor and demand a halt to the rebel support if they want to stem the migration tide. Alternatively, the chaos of a civil conflict could cause those refugees to flee without the government's intervention. In both scenarios, the state faced with a new, unwanted, refugee population has the same choice: If they continue to support the rebel group, they will continue to receive refugee in-flows. If they stop supporting the rebel group, the refugee in-flow will subside either because the government stops forcing them out or because they no longer have the incentive to leave when the government subdues the rebellion. The only difference is the coercive threat, and we expect a reactance response in that condition. Indeed, this example is inspired by Greenhill's (2010) anomalous finding about coercive engineered migration's high rate of success. If the target of coercion does not perceive their challenger to intentionally threaten their freedom of choice, reactance does not apply.[19]

5 Study 1: Any Disputed Island in a Storm

Study 1 was a 3 group, between-subjects experiment fielded to a sample of 590 participants on Amazon's Mechanical Turk in August 2019. Participants, 50.68% of whom identified as male and 70.85% as White/Caucasian were located in the U.S. Mechanical Turk samples are not representative of the broader U.S. population but are increasingly common in political science (e.g., Tomz and Weeks, 2013; Huff and Kertzer, 2018; Kertzer, Renshon and Yarhi-Milo, 2019) and more diverse

[19]Indeed, Greenhill (2010, 60) notes that "few of these [migration] crises ever reach the desk of target state executive(s)." We nevertheless refrain from using this example in our experimental designs because doing so requires the tenuous assumption that all participants believe that refugee inflows are unambiguously costly.

than other convenience samples (Berinsky, Huber and Lenz, 2012; Huff and Tingley, 2015).[20]

5.1 Methods and Materials

The experiment proceeded in four steps. First, all participants completed a standard battery of pre-treatment questions to measure partisanship, ideology, and militarism (Herrmann, Tetlock and Visser, 1999). Second, participants received an introduction explaining that they were about to read about a generic scenario that might take place in international politics in the future, but is not meant to represent any particular countries or conflicts (following, e.g., Tomz, 2007; Tomz and Weeks, 2018; Kertzer, Renshon and Yarhi-Milo, 2019). This was followed by an overview of the conflict: The participant's country withdrew forces stationed on a disputed island due to an incoming hurricane. They then learned that Navalia — a neighboring state who also claims the island — is attempting to retake the territory while it remains unguarded. The scenario details, included below, hold constant the two countries' relative power and conflict history:

> Here is the situation:
>
> - Your country is involved in a dispute with the country of Navalia over a small island. Both your country and Navalia claim to own the island, which has valuable resources.
> - Both countries are equally powerful.
> - Your country is not at war with Navalia and has never fought Navalia before.
> - Your country has stationed soldiers on the island for many years. They were withdrawn temporarily due to an incoming hurricane.
> - You recently ordered a small supply ship to return your countrys troops to the island.
> - However, a spy for your country just learned that Navalia is sending a military force to seize the island before your troops can return.
> - Their orders are to seize the island if they arrive first but to turn back if your troops have already returned to defend it.

Third, participants were randomly assigned to one of three treatment groups. In the *natural costs* condition, the ship must travel through the hurricane to reach the island in time, and there is

[20] In line with current best practices, we required workers to have completed > 100 previous tasks with a 95% approval rating and to be located in the U.S. We compensated participants $1.40 for a task that took an average of 10.1 minutes to complete. The study was approved by IRBs at both Dartmouth College and Georgia State University. Following Burleigh, Kennedy and Clifford's (2018) recommendations, we screened out a small number of participants (< 20) using foreign VPSs. The sample size is 590 after excluding these participants.

a chance that the ship will sink and cause 100 soldiers to die as a result. In the *coercion* condition — indicated by brackets in the text below — a submarine captain from Navalia issues a direct threat. If the respondent's country's ship does not turn back, they will try to sink it and kill all 100 soldiers on board.[21] Notably, and in line with our empirical strategy, the only difference between these two treatments is the existence of a coercive threat. In both situations, participants face the same cost/benefit trade-off: order the ship to continue and face a 50% chance that the soldiers die but a 50% chance that you retain the valuable island, or order the ship to turn around and take the certain territorial loss but keep the ship and soldiers on board. In a third, no-information *control* condition, participants do not receive any information about potential costs from turning back.

The abstract nature of this scenario allows us to control for reputation and audience costs by providing fixed values. All participants, including the control group, receive the final two pieces of information listed above. They are told that the public is not aware of the dispute (Kertzer, Renshon and Yarhi-Milo, 2019) to control for audience costs. While reputation has been conceptualized in too many ways to list here (see Brutger and Kertzer, 2018 for a review), this study focuses on concerns about future predation and informs participants that they are unlikely to face this challenger again.[22]

- To reach the island before Navalia's forces, your supply ship must travel **through the hurricane.** [*past one of Navalia's submarines.*
- *The captain of that submarine just got on the radio and* ***demanded that your ship turn back.***
- *He threatened you:* ***"Turn back now, or else we will fire."***]
- If you order your ship to continue on, your best guess is that it has a 50% chance of surviving to reach the islands and a 50% chance of being sunk.
- If the ship sinks, **all 100 soldiers on board will die.**
- Members of the public are not aware of this dispute due to its isolated location.
- This situation is unusual, and it is unlikely to arise again since your country and Navalia do not interact frequently.

Fourth, we measured our key dependent variables and proposed mechanisms. Participants re-

[21] We verified that this was considered a costly outcome in a pre-test where we asked participants to record the extent to which they agree or disagree that their country should go to great lengths to avoid 100 military deaths. On a 7-point scale, the mean was 5.8, suggesting that most people believe that this is a high cost to pay.

[22] We include one factual manipulation check that asked participants whether, in the scenario they read, someone from Navalia directly threatened them. In the coercive threat condition, 91% of respondents correctly selected yes, and 89% of respondents in the both the control and natural consequences condition correctly selected no.

sponded to a binary question asking them how they would respond to the situation — order the ship to return or to continue on to the island. To assess whether coercive threats provoke aggression (hypothesis 2), we asked participants whether they would support or oppose a) economic sanctions or b) war with Navalia in response to the situation (5-point scale rescaled from 0 to 1). To measure anger, we create a 4-item scale based on the extent to which participants report feeling angry, furious, irritated, and annoyed with Navalia (rescaled from 0 to 1; $\alpha = 0.91$) (Quick and Stephenson, 2007; Valentino et al., 2011; Gadarian and Albertson, 2014).[23] We assess perceived threat to freedom of choice with an item that asks whether participants agree that "someone from Navalia is trying to manipulate me." Finally, a 100-point feeling thermometer taps whether participants feel cold toward Navalia (rescaled 0 to 1). The survey concludes with a 6-item knowledge scale and demographic questions.

5.2 Results

We present our results in four stages. First, we present average treatment effects for the three primary dependent variables to show that people who receive a coercive threat are less likely to capitulate, more likely to support going to war with their opponent, and more likely to support sanctioning their opponent, relative to those whose freedom is threatened by an impersonal force. Second, we analyze the effect of our treatments on indicators associated with reactance, including anger and evaluations of Navalia, and altenative explanations. We find that coercion increases anger and perceived manipulation. While we find no evidence that our treatments affected concerns for audience costs, coercive threats raise reputation concerns. In the third section, we estimate causal mediation modest to find that reputation concerns are responsible for part of coercion's effect, but that the choice to stand firm is primarily mediated by perceived manipulation.

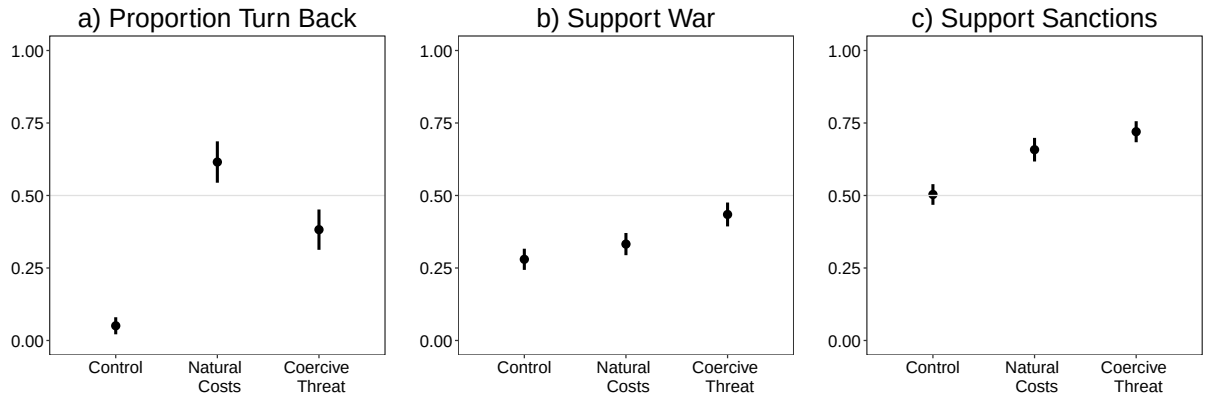
Coercion Failure

Figure 1 displays cell means for our three primary dependent variables and 95% confidence intervals, by treatment group. Turning first to panel **a**, we find strong support for hypothesis 1 — when participants receive a direct and personal threat, they are more likely to stand firm compared to those for whom recalcitrance poses the same risks but who do not receive a threat. In the coercion

[23] We include 4 distractor items — anxious, mournful, sad, and excited — to ameliorate acquiescence bias. The eight target emotions are presented in random order.

group, 61.5% of respondents chose to continue on in hopes of retaining their island, compared to 38.2% in the natural consequences condition ($p < 0.01$). Both treatments were much more likely to turn back compared to the baseline control condition, where only 5% of participants were willing to abandon their claim to the island when there were no apparent costs associated with trying to retain it.

Figure 1: Study 1: Response to Threat by Treatment



Note: N=590. Figure displays the average score on each dependent variable and 95% confidence intervals. Panel **a** displays proportion of respondents who chose to "turn back", and panels **b** and **c** display average support for war with Navalia and opposing sanctions, respectively.

The results in panel **a** show that targets of coercive threats are less likely to give up their territory (H1), but panels **b** and **c** demonstrate that they are also more likely to support an aggressive policy stance against their challenger (H2). On average, receiving a coercive threat increases respondents' support for war by 0.15 points on the 0 to 1 scale compared to the control group ($p < 0.01$), and 0.1 compared to the natural costs group ($p < 0.01$). Those who face a hurricane rather than an attack submarine, however, are not significantly more likely than those in the control group to support going to war with Navalia ($p = 0.06$).

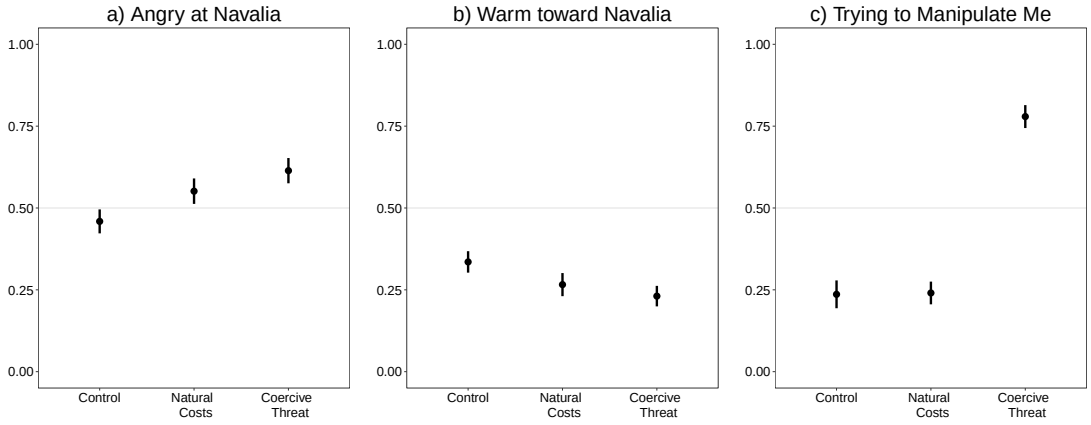
Imposing sanctions against Navalia for their attempt to seize the island is a conflictual response but less aggressive than going to war (Goldstein, 1992), and panel **c** shows that all groups were at least somewhat amenable to punishing their opponent for targeting their territorial possession. Exposure to a dangerous hurricane increased support for sanctions compared to the control group ($p < 0.01$). Even respondents whose path is impeded by an impersonal force are willing to punish Navalia for putting them in this predicament as long as it does not entail using force. Consistent with our expectations, the threat treatment increased support for sanctions compared to both the

control ($p < 0.01$) and natural costs treatment ($p < 0.025$).

Reactance Indicators

Next, we turn to three observable outcomes associated with a a reactance-based resistance to threat. People respond to threats to their freedom by performing the prohibited action or by becoming aggressive against the source of the threat, and we find evidence to support these expectations in the section above (Dillard and Shen, 2005). A complete test of our theory, however, requires evidence that the coercive threat treatment affects the emotional and cognitive outcomes that constitute reactance. The threat should cause participants to feel that they are being manipulated at the same time that it inspires anger and derogation of the challenger (H3 and H4). To test these expectations, we analyze the average differences across treatment groups in self-reported anger, feelings of warmth toward Navalia, and respondent beliefs that someone is trying to manipulate them. Panels **a**, **b** and **c** in Figure 2 display cell means for these three outcomes.

Figure 2: Study 1: Reactance Mechanisms



Note: N=590. Figures display the average score on each dependent variable and 95% confidence intervals by experimental condition. Higher values in panel represent **a** represent greater anger, **b** warmer ratings of Navalia, and **c** agreement that Navalia is trying to manipulate you.

Compared to both the control ($p < 0.01$) and natural costs treatment groups ($p < 0.05$), participants who received a coercive threat reported that they felt angrier toward Navalia. The coercive threat increased anger by 0.063 on a 0 to 1 scale compared to the hurricane. This provides support for Hypothesis 3 and additional evidence for our psychological theory of resistance to threats. While threatened respondents are angry respondents, the feeling thermometer results reveal no statistically

significant differences in feelings of warmth toward the challenger ($p = 0.15$). Respondents in both treatment groups feel cooler toward Navalia than their counterparts in the control condition (both $p < 0.01$).

The results in panel **c** show that the presence of a coercive threat has a dramatic effect on whether participants believe that their opponent is trying to manipulate them. Although Navalia is attempting to seize their territory on a technicality in all three treatment groups — the participant's soldiers retreated from the island temporarily to wait out the storm and Navalia saw the opportunity to simply move in — only respondents subject to a threat feel that they are being manipulated (both $p < 0.01$). Consistent with psychological reactance, they express concern that a personal force is trying to direct their free behavior.

Alternative Explanations and Causal Mediation

The research design controls for several alternative explanations at once — we hold constant the costs of capitulation, the size of the demand, credibility,[24] and the challenger's power,[25] for example. We also sought to control for audience costs and reputation, two important alternative mechanisms, by fixing their values in the vignette — we tell participants that the dispute is in an isolated area and that the outcome of this dispute is unlikely to have implications for their future conflicts with Navalia. To test whether participants nevertheless drew inferences that implicated these factors, we include two items that ask participants to rate the probability that a) their public will disapprove of them as a leader and b) their country's reputation will suffer if they turn back the supply ship and Navalia takes control of the island.

Model 1 in Table 1 shows that neither treatment group significantly increased respondents' expectations that the public would take exception to their decision to turn back. The difference between the two treatment groups is similarly non-significant ($p = 0.09$). Non-significant coeffi-

[24] To verify that the treatments did not affect participant estimates of the chance that they would avoid consequences by turning back, we asked them to estimate on a 5-point scale the chance that their country's soldiers would die if they turned back. A series of pairwise comparisons shows that there are no significant differences between either treatment and the control, or between the two experimental treatments.

[25] One potential concern is that participants could change their perception of the challenger's power based on the presence or absence of a threat. If participants believe that issuing a threat reveals a challenger's weakness, because they have threatened them rather than simply taken the island by fait accompli, they might be more likely to stand firm and this would confound our results. To test whether the treatments had an effect on information about the challenger's power, we measure participant evaluations of how strong Navalia is, relative to their country on a 5-point scale from "vastly weaker" to "vastly stronger" post-treatment. The mean response for the full sample is 2.8, slightly below the midpoint value of about equal in power, and this does not vary by experimental condition. A series of pairwise difference-in-means tests reveals that mean power estimates are roughly the same across each group, from a low of 2.79 in the coercion treatment to a high of 2.82 in the control group.

Table 1: Study 1: Reputation and Audience Costs

	Public Disapprove	Reputation Suffer
	(1)	(2)
Natural Costs	−0.027	−0.041
	(0.027)	(0.030)
Coercion	0.021	0.026
	(0.027)	(0.029)
Constant	0.539**	0.544**
	(0.018)	(0.020)
N	590	590
R^2	0.005	0.008

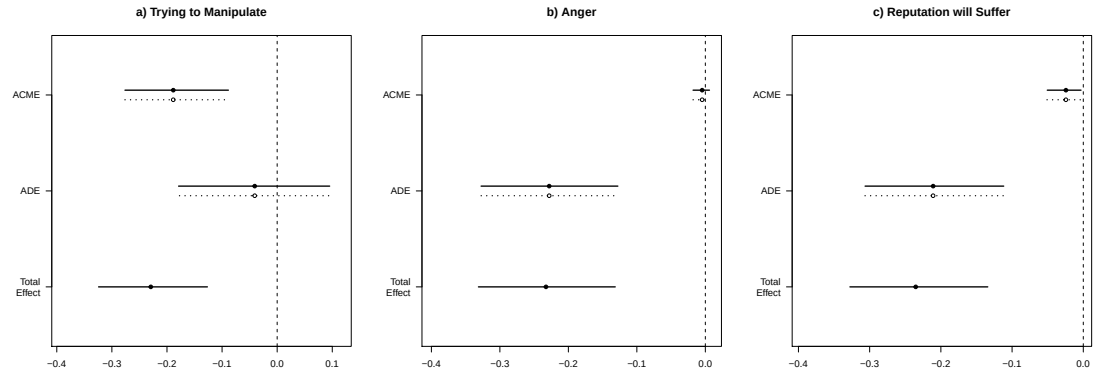
*p < .05; **p < .01
Note: Table displays OLS coefficients, dependent variables are rescaled to range from 0 to 1. Higher values indicate perceptions that this outcome "definitely will" happen. The control condition is the reference category.

cients on the treatment conditions in Model 2 suggest that neither treatment increased reputation cost perceptions relative to the control. Yet if we compare the two treatment groups, we find that participants who received a coercive threat more likely to say that their country's reputation would suffer if they ceded the island to Navalia ($p < 0.05$), compared to those whose decision was constrained by the weather. Backing down could threaten a country's reputation for resolve and thus give participants an incentive to push forward in spite of short-term costs (Brutger and Kertzer, 2018; Kertzer, Renshon and Yarhi-Milo, 2019; Dafoe, Zhang and Caughey, 2018). These results imply that coercive threats inflate reputation concerns, but not domestic audience costs, even when we attempt to control for them experimentally.

To evaluate the extent to which the effect of our coercion treatment on participants' decision to turn back or stand firm is mediated by reputation and psychological reactance,[26] we estimate a series of non-parametric causal mediation models and plot the results in Figure 3, focusing on the comparison between the two costly treatment groups. The plots include three quantities of interest for each of three prospective mediators — perceived manipulation, anger, and reputation concerns. The average causal mediation effect (ACME) shows the effect of the coercion treatment that flows through the mediator, and the average direct effect (ADE) is the treatment effect as channeled through all other mechanisms. The total effect is the sum of the ACME and ADE. If our theory

[26]Because we did not find evidence for direct effects of coercion on audience costs and the Navalia feeling thermometer, we do not estimate mediation models for them. We focus on the three variables that showed a signficant difference between the two treatments.

Figure 3: Study 1: Reactance Mechanisms



Note: N=590. Figures display the Average Causal Mediation Effects (ACME), Average Direct Effects (ADE), and Total Effects for the difference between the Coercion and Natural Consequences treatments from a series of non-parametric mediation analyses using the mediation package in R (Imai et al., 2011) with 95% quasi-Bayesian confidence intervals. The binary dependent variable is coded 1 if participants chose to turn back, and 0 if they chose to continue forward. Mediators have all been rescaled to range from 0 to 1, and higher values represent more agreement that they are being manipulated, anger, and belief that the country's reputation will suffer. Solid points and lines plot quantities for the Coercion group, and dashed line represents the natural consequences group. Estimates differ slightly due to the non-linear dependent variable.

is correct, a significant portion of the difference between the two treatment conditions should be mediated by anger and participants' belief that they are being manipulated — that their freedom to choose is under siege.

The results reveal that the effect of the coercive threat relative to a costly impersonal force, is primarily mediated by the degree to which participants believe that they are being manipulated. Panel **a** in Figure 2 shows that this mechanism accounts for 83% of the total effect of the coercion treatment, whereas panel **c** shows that 10% of the observed difference can be attributed to beliefs that the country's reputation will suffer in the eyes of other countries. Coercive threats give rise to reputation concerns which in turn drive resistance, but it is responsible for a smaller share of coercion's total effect. Against our expectations, we do not find evidence that anger significantly mediates the relationship between coercive threats and resistance to capitulation.

6 Study 2: Terrorist Safe Havens

In the second planned experiment, respondents are confronted with a hypothetical scenario featuring a terrorist group that is setting up bases in Eritrea. Some respondents are told that the Eritrean dictator threatened them, indicating he would allow the terrorist bases to remain unless the U.S.

halts support for a nonviolent pro-democracy movement opposing his rule. Other respondents receive a version of the scenario without any Eritrean threats or demands. They are told that the dictator is using all of his capabilities to subjugate the pro-democracy movement. Weakening that movement by withdrawing support would free up the resources for the dictator to take action and destroy the terrorist bases. Study 1 showed the relevance of reputation to coercion failure. Study 2 allows us to more plausibly minimize respondent concerns about reputation and to directly test the impact of making them salient. However, those advantages come at the expense of greater complexity, which increases the risk that respondents will react to something other than the treatment or misunderstand the scenario.

6.1 Methods and Materials

Study 2 follows the same steps as Study 1, with the exception of an additional randomization relating to reputation and audience costs. Respondents are placed in the role of U.S. Ambassador to Eritrea and asked to respond to a hypothetical situation they might encounter. As presented below, a terrorist group is setting up bases in Eritrea, posing a threat to the United States. Meanwhile, the United States has been secretly furnishing aid to a pro-democracy movement challenging the Eritrean dictatorship through nonviolent protest. Respondents will need to choose between these two objectives.

Here is the situation:

- A terrorist group affiliated with ISIS has sworn to attack the United States.
- This terrorist group is setting up bases in the country of Eritrea.
- Because U.S. intelligence agencies have been unable discover the exact locations of these bases, **you need the cooperation of the Eritrean government to remove them.**
- The Eritrean dictator has the ability to destroy the terrorist bases.
- Eritrea is ruled by an oppressive dictator who is focused on cracking down on a pro-democracy movement that is challenging him with nonviolent protests.
- The United States is secretly aiding that Eritrean pro-democracy movement.

Respondents are randomly assigned to either a *coercion* condition or a *natural costs* condition. In the coercion condition, the dictator demands that the U.S. cease support for the pro-democracy

movement and threatens to permit the terrorists to operate from Eritrea if that demand goes unmet. In the natural costs condition, respondents are told that cracking down on the pro-democracy movement is exhausting dictator's capacity to deal with internal threats. Although the dictator has made no threat or demand, easing the threat from the movement would free the resources needed to remove the terrorist bases. In both conditions, respondents read that allowing the terrorist bases to go unchecked will likely cost hundreds of American lives. Below, the italicized text containing a verbal threat replaces the first bullet point from the natural consequences condition to create the coercion condition; this is the only difference between the two vignettes.

- Although the dictator hasn't said anything to you, you know that **the pro-democracy movement is using up all of Eritrea's resources, so there are none left to deal with the terrorist bases.**
- *In a private meeting earlier today,* ***the dictator threatened you: "You must stop meddling with Eritrea, or else!"***
- If you have the U.S. withdraw support for the pro-democracy movement, the Eritrean government would be able to conduct a series of raids to eliminate the terrorist bases.
- The CIA assesses that, if able to operate from these bases, the terrorist group will be able to conduct attacks in the United States in the next year, likely **killing hundreds of Americans.**
- The CIA is confident that the Eritrean government would remove the bases if the pro-democracy movement weakens.
- The CIA assesses that there is a **low likelihood of future conflicts** with Eritrea.
- Ending support for the pro-democracy movement would happen **secretly**. It would be unlikely to receive global attention.

To further assess reputation and audience costs as alternative explanations for coercion failure, we randomly assign participants to receive either a treatment that minimizes both possible mechanisms or one that makes both salient. We manipulate these two factors jointly because our primary interest lies in evaluating whether participants will resist coercion when both are definitively absent, not in evaluating the two explanations against one another; we leave that task to other research (Dafoe, Hatz and Zhang, 2017). The text above shows the planned language for the low-salience condition, and participants in the high-salience condition receive the following two points instead:

- The CIA assesses that there is a **high likelihood of future conflicts** with Eritrea.
- Ending support for the pro-democracy movement would happen **publicly**. It would be likely to receive global attention.

Although concessions made privately are hidden from both the international community and domestic audiences, the adversary receiving the concession still necessarily knows about it. Unable to eliminate that potential for the Eritrean dictator's perception of the U.S. to change (i.e., the U.S. reputation with the Eritrean dictator), we instead minimized its importance by downplaying the likelihood of future conflicts between the United States and Eritrea. Choosing a small, distant state such as Eritrea with little history of conflict with the United States helped make this possible. We expect that few respondents will have knowledge of Eritrea. They may recognize it as a real country of relatively small size that rarely appears in U.S. news reports.[27] We also selected Eritrea due to its entrenched dictatorship and proximity to terrorist groups in Somalia and Yemen.

Respondents then answer a series of questions designed to measure the dependent variable — coercion failure — and assess other observable implications of reactance. Most importantly, they first answer the question shown below. Per Hypothesis 1, we expect respondents to more frequently continue to support the pro-democracy movement in the coercion condition than the natural consequences condition. Such a result would offer evidence that reactance is causing coercion failure.

What do you think the U.S. should do in this situation?

- The U.S. should increase its support for the pro-democracy movement opposing the Eritrean dictator.
- The U.S. should continue to provide about as much support to the pro-democracy movement opposing the Eritrean dictator as it currently does.
- The U.S. should continue to provide support to the pro-democracy movement opposing the Eritrean dictator, but only occasionally and less than current levels.
- The U.S. should halt all support for the pro-democracy movement opposing the Eritrean dictator.

[27] In a pretest, many participants offered in a comment section that they had never heard of Eritreat or mused about whether it was a real place.

Respondents then indicate their degree of support for sanctioning Eritrea, military strikes against Eritrea, and invading Eritrea to take control of the country. Per Hypothesis 2, we expect to observe a greater willingness to use sanctions and force against Eritrea in the coercion condition. Finally, respondents complete the anger scale and feeling thermometers designed to gauge their attitudes and emotions toward the Eritrean Government and the pro-democracy movement. Hypotheses 3-4 predict that respondents will harbor more negative emotions toward and views about the Eritrean Government in the coercion condition.

6.2 Results

We look forward to your feedback on the design described above, so that one day this section will include more text.

7 Conclusion

This study investigated the possibility that a psychological constraint against capitulation to coercive threats could be a powerful force in international politics. Drawing on literature from psychology, we put forward reactance as a potential explanation for the mounting evidence that most types of coercion fail most of the time. Indeed, reactance can be the answer to what we describe as the puzzle of coercion failure: the tendency of coercion to fail with such surprising frequency, contrary to standard rational-actor models. Rather than coldly calculate costs, benefits, and credibility when deciding whether to concede, we suspect that leaders dig in their heels, refuse to relent, and seek opportunities to push back against their challenger.

To assess this possibility, we developed two experiments that each exploited distinctive, unusual circumstances in which plausible "natural consequences" mimic a coercive threat in every way other than the presence of the threat itself. In Study 1, 60 percent of respondents chose to defy a submarine captain's threat to sink their supply ship if it attempts to deploy troops to keep control of a disputed island. When faced with the same options, costs, and benefits but with a storm as the source of the risk (rather than coercion), only 38 percent chose to defy the storm to retain the island. The results further reveal increased anger in the coercion condition and an increased willingness to sanction the coercer. Causal mediation analysis suggests that reputation incentives were responsible for a small portion of the difference in capitulation rates between the coercion and natural costs treatment

groups.

Our use of the experimental method makes it possible to draw uniquely strong causal inferences. While we acknowledge that results from a survey experiment do not automatically generalize to the behavior of leaders, we believe that establishing the tendency for large samples of individuals to respond in the predicted manner is a necessary and informative first step toward applying the concept of reactance to state behavior, and mounting evidence suggests that elites respond to experimental treatments much like ordinary citizens (Kertzer, Renshon and Yarhi-Milo, 2019). Moreover, the primary factors that might moderate the effect of reactance, such as a person's belief that she possesses freedom-of-choice in a particular domain or her objective power (Miron and Brehm, 2006), should be more prominent, not less prominent, among heads of state (Renshon, 2017). This would suggest that our average effect estimates based on a convenience sample of the regular population are too conservative rather than improbably large.

The potential implications are far-reaching. If psychology (reactance) is a potent cause of coercion failure, then it stands to reason that it also causes war. The bargaining model of war, perhaps the leading theoretical framework for understanding the causes of war, expects that war breaks out when states cannot reach a deal to avoid it.[28] Due to the costliness of war, such a bargain should frequently be available; why not simply agree to the likely eventual outcome of the war while bypassing the costs of fighting to reach it? The framework envisions that the dissatisfied state will demand concessions and threaten war, avoiding war when a credible threat to start it results in peaceful concessions (coercive bargaining). If human psychology interferes with this process because reactance impedes concessions, war could all too easily be the result.

Losing faith in coercive diplomacy could change how leaders approach all manner of foreign policy problems. They might forgo coercive threats altogether if they see them as futile, or attempt to overcome reactance by threatening ever-higher costs. They might instead try to impose desired outcomes rather than rely on demands and threats.[29] Of course, leaders inclined to adopt more aggressive strategies often take this approach regardless — after their threat fails to achieve its desired end. Understanding the full extent of the obstacles to coercive success might lead states to skip inefficient periods of attempted coercion in favor of brute force.In the end, futile threats benefit neither coercers nor targets.

[28] Fearon (1995) develops rationalist explanations for bargaining failure and thus war, but explicitly does not evaluate psychological explanations.

[29] On this possibility, see Schelling (1966); Altman (2017).

References

Abrahms, Max. 2006. "Why Terrorism Does Not Work." *International Security* 31(2):42–78.

Altman, Dan. 2017. "By Fait Accompli, Not Coercion: How States Wrest Territory from Their Adversaries." *International Studies Quarterly* 61(4):881–891.

Arreguin-Toft, Ivan. 2001. "How the Weak Win Wars: A Theory of Asymmetric Conflict." *International security* 26(1):93–128.

Art, Robert J and Kelly M Greenhill. 2018. "The Power and Limits of Compellence: A Research Note." *Political Science Quarterly* 133(1):77–98.

Ashworth, Scott, Joshua D Clinton, Adam Meirowitz and Kristopher W Ramsay. 2008. "Design, Inference, and the Strategic Logic of Suicide Terrorism." *American Political Science Review* 102(2):269–273.

Bapat, Navin A and T Clifton Morgan. 2009. "Multilateral versus Unilateral Sanctions Reconsidered: A Test Using New Data." *International Studies Quarterly* 53(4):1075–1094.

Behrouzian, Golnoosh, Erik C Nisbet, Aysenur Dal and Ali Çarkoğlu. 2016. "Resisting censorship: How citizens navigate closed media environments." *International Journal of Communication* 10:23.

Berejikian, Jeffrey D. 2002*a*. "A cognitive theory of deterrence." *Journal of Peace Research* 39(2):165–183.

Berejikian, Jeffrey D. 2002*b*. "Model building with prospect theory: A cognitive approach to international relations." *Political Psychology* 23(4):759–786.

Berinsky, Adam J, Gregory A Huber and Gabriel S Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political analysis* 20(3):351–368.

Betts, Richard K. 2010. *Nuclear Blackmail and Nuclear Balance*. Brookings Institution Press.

Brehm, Jack W. 1966. *A theory of psychological reactance*. New York: Academic Press.

Brehm, Sharon S and Jack W Brehm. 2013. *Psychological reactance: A theory of freedom and control*. Academic Press.

Brutger, Ryan and Joshua D Kertzer. 2018. "A dispositional theory of reputation costs." *International Organization* 72(3):693–724.

Burleigh, Tyler, Ryan Kennedy and Scott Clifford. 2018. "How to screen out VPS and international respondents using Qualtrics: A protocol." *Available at SSRN* .

Cherulnik, Paul D and Murray M Citrin. 1974. "Individual difference in psychological reactance: The

interaction between locus of control and mode of elimination of freedom." *Journal of Personality and Social Psychology* 29(3):398.

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. "Information equivalence in survey experiments." *Political Analysis* 26(4):399–416.

Dafoe, Allan, Jonathan Renshon and Paul Huth. 2014. "Reputation and Status as Motives for War." *Annual Review of Political Science* 17:371–393.

Dafoe, Allan, Sophia Hatz and Baobao Zhang. 2017. "Coercion and provocation." *Unpublished working paper. http://www. allandafoe. com/provocation* .

Dillard, James Price and Lijiang Shen. 2005. "On the nature of reactance and its role in persuasive health communication." *Communication Monographs* 72(2):144–168.

Downes, Alexander B and Lindsey A O'Rourke. 2016. "You Can't Always Get What You Want: Why Foreign-Imposed Regime Change Seldom Improves Interstate Relations." *International Security* 41(2):43–89.

Drezner, Daniel W. 1999. *The Sanctions Paradox: Economic Statecraft and International Relations*. Cambridge University Press.

Drezner, Daniel W. 2011. "Sanctions sometimes smart: targeted sanctions in theory and practice." *International Studies Review* 13(1):96–108.

Duelfer, Charles A and Stephen Benedict Dyson. 2011. "Chronic misperception and international conflict: The US-Iraq experience." *International Security* 36(1):73–100.

Early, Bryan R. 2015. *Busted Sanctions: Explaining Why Economic Sanctions Fail*. Stanford University Press.

Erceg-Hurn, David M and Lyndall G Steed. 2011. "Does exposure to cigarette health warnings elicit psychological reactance in smokers?" *Journal of Applied Social Psychology* 41(1):219–237.

Fearon, James. 2002. "Selection Effects and Deterrence." *International Interactions* 28(1):5–29.

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):379–414.

Feaver, Peter D and Eric B Lorber. 2015. "The sanctions myth." *The National Interest* (138):22–27.

Fortna, Virginia Page. 2015. "Do Terrorists Win? Rebels' Use of Terrorism and Civil War Outcomes." *International Organization* 69(3):519–556.

Gadarian, Shana Kushner. 2014. "Beyond the waters edge: Threat, partisanship, and media." *The political psychology of terrorism fears* pp. 67–84.

Gadarian, Shana Kushner and Bethany Albertson. 2014. "Anxiety, immigration, and the search for information." *Political Psychology* 35(2):133–164.

Gartzke, Erik. 2013. "The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth." *International Security* 38(2):41–73.

Goemans, Hein Erich. 2000. *War and Punishment: The Causes of War Termination and the First World War*. Princeton University Press.

Goldstein, Joshua S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution* 36(2):369–385.

Greenhill, Kelly M. 2010. *Weapons of Mass Migration: Forced Displacement, Coercion, and Foreign Policy*. Cornell University Press.

Hall, Todd H. 2011. "We will not swallow this bitter fruit: Theorizing a diplomacy of anger." *Security Studies* 20(4):521–555.

Hall, Todd H. 2017. "On Provocation: Outrage, International Relations, and the Franco–Prussian War." *Security Studies* 26(1):1–29.

Haun, Phil. 2015. *Coercion, Survival, and War: Why Weak States Resist the United States*. Stanford University Press.

Herrmann, Richard. 1994. "Coercive diplomacy and the crisis over Kuwait, 1990-1991." *The limits of coercive diplomacy* pp. 229–264.

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

Horowitz, Michael and Dan Reiter. 2001. "When Does Aerial Bombing Work? Quantitative Empirical Tests, 1917-1999." *Journal of Conflict Resolution* 45(2):147–173.

Hufbauer, Gary Clyde, Jeffrey J Schott and Kimberly Ann Elliott. 1990. *Economic Sanctions Reconsidered: History and Current Policy*. Vol. 1 Peterson Institute.

Huff, Connor and Dustin Tingley. 2015. "Who are these people? Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research & Politics* 2(3):2053168015604648.

Huff, Connor and Joshua D Kertzer. 2018. "How the public defines terrorism." *American Journal of Political Science* 62(1):55–71.

Hyde, Susan D. 2015. "Experiments in international relations: Lab, survey, and field." *Annual Review of Political Science* 18:403–424.

Imai, Kosuke, Luke Keele, Dustin H. Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765–789.

Jervis, Robert. 1992. "Political implications of loss aversion." *Political psychology* pp. 187–204.

Jones, Seth G and Martin C Libicki. 2008. *How Terrorist Groups End: Lessons for Countering Al Qa'ida*. Vol. 741 Rand Corporation.

Katz, Sherri Jean, Sahara Byrne and Alyssa Irene Kent. 2017. "Mitigating the perception of threat to freedom through abstraction and distance." *Communication Research* 44(7):1046–1069.

Kertzer, Joshua D and Dustin Tingley. 2018. "Political psychology in international relations: beyond the paradigms." *Annual Review of Political Science* 21:319–339.

Kertzer, Joshua D, Jonathan Renshon and Keren Yarhi-Milo. 2019. "How Do Observers Assess Resolve?" *British Journal of Political Science* pp. 1–23.

Krasner, Stephen D. 1999. *Sovereignty: organized hypocrisy*. Princeton University Press.

Krause, Peter. 2013. "The Political Effectiveness of Non-State Violence: A Two-Level Framework to Transform a Deceptive Debate." *Security Studies* 22(2):259–294.

Kroenig, Matthew. 2018. *The Logic of American Nuclear Strategy: Why Strategic Superiority Matters*. Oxford University Press.

Kydd, Andrew H and Barbara F Walter. 2006. "The strategies of terrorism." *International security* 31(1):49–80.

LaFree, Gary, Laura Dugan and Raven Korte. 2009. "The impact of British counterterrorist strategies on political violence in Northern Ireland: Comparing deterrence and backlash models." *Criminology* 47(1):17–45.

Laurin, Kristin, Aaron C Kay, Devon Proudfoot and Gavan J Fitzsimons. 2013. "Response to restrictive policies: Reconciling system justification and psychological reactance." *Organizational Behavior and Human Decision Processes* 122(2):152–162.

LaVoie, Nicole R, Brian L Quick, Julius M Riles and Natalie J Lambert. 2017. "Are graphic cigarette warning labels an effective message strategy? A test of psychological reactance theory and source appraisal." *Communication Research* 44(3):416–436.

Lindsay, Jon R. 2013. "Stuxnet and the Limits of Cyber Warfare." *Security Studies* 22(3):365–404.

Mack, Andrew. 1975. "Why Big Nations Lose Small Wars: The Politics of Asymmetric Conflict." *World politics* 27(2):175–200.

Marwicka, Robin. 2018. *Emotional choices: How the logic of affect shapes coercive diplomacy*. Oxford University Press.

McDermott, Rose. 2017. "The evolved psychology of coercion." *Comparative Strategy* 36(2):91–98.

Mercer, Jonathan. 2010. *Reputation and International Politics*. Cornell University Press.

Miller, Claude H, Bobi Ivanov, Jeanetta Sims, Josh Compton, Kylie J Harrison, Kimberly A Parker, James L Parker and Joshua M Averbeck. 2013. "Boosting the potency of resistance: Combining the motivational forces of inoculation and psychological reactance." *Human Communication Research* 39(1):127–155.

Miron, Anca M and Jack W Brehm. 2006. "Reactance theory-40 years later." *Zeitschrift für Sozialpsychologie* 37(1):9–18.

Moore, Celia and Lamar Pierce. 2016. "Reactance to transgressors: why authorities deliver harsher penalties when the social context elicits expectations of leniency." *Frontiers in psychology* 7:550.

Morgan, T Clifton, Navin Bapat and Yoshiharu Kobayashi. 2014. "Threat and Imposition of Economic Sanctions 1945-2005: Updating the TIES Dataset." *Conflict Management and Peace Science* 31(5):541–558.

Nisbet, Erik C, Kathryn E Cooper and R Kelly Garrett. 2015. "The partisan brain: How dissonant science messages lead conservatives and liberals to (dis) trust science." *The ANNALS of the American Academy of Political and Social Science* 658(1):36–66.

Nooruddin, Irfan. 2002. "Modeling selection bias in studies of sanctions efficacy." *International Interactions* 28(1):59–75.

Nyhan, Brendan and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32(2):303–330.

O'Rourke, Lindsey A. 2018. *Covert Regime Change: America's Secret Cold War*. Cornell University Press.

Pape, Robert A. 1996. *Bombing to Win: Air Power and Coercion in War*. Cornell University Press.

Pape, Robert A. 1997. "Why Economic Sanctions Do Not Work." *International security* 22(2):90–136.

Pape, Robert A. 2003. "The Strategic Logic of Suicide Terrorism." *American political science review* 97(3):343–361.

Peffley, Mark and Jon Hurwitz. 2007. "Persuasion and resistance: Race and the death penalty in America." *American Journal of Political Science* 51(4):996–1012.

Powell, Robert. 1990. *Nuclear Deterrence Theory: The Search for Credibility*. Cambridge University Press.

Press, Daryl Grayson. 2005. *Calculating Credibility: How Leaders Assess Military threats*. Cornell University Press.

Quick, Brian L and Jennifer R Considine. 2008. "Examining the use of forceful language when designing exercise persuasive messages for adults: A test of conceptualizing reactance arousal as a two-step process." *Health communication* 23(5):483–491.

Quick, Brian L and Michael T Stephenson. 2007. "Further evidence that psychological reactance can be modeled as a combination of anger and negative cognitions." *Communication Research* 34(3):255–276.

Reiter, Dan. 2009. *How Wars End*. Princeton University Press.

Renshon, Jonathan. 2017. *Fighting for status: Hierarchy and conflict in world politics*. Princeton University Press.

Schelling, Thomas C. 1966. *Arms and Influence*. Vol. 190 Yale University Press.

Sechser, Todd S. 2010. "Goliath's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64(4):627–660.

Sechser, Todd S. 2011. "Militarized Compellent Threats, 1918-2001." *Conflict Management and Peace Science* 28(4):377–401.

Sechser, Todd S and Matthew Fuhrmann. 2017. *Nuclear Weapons and Coercive Diplomacy*. Cambridge University Press.

Slantchev, Branislav L. 2005. "Military Coercion in Interstate Crises." *American Political Science Review* 99(4):533–547.

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.

Steindl, Christina, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch and Jeff Greenberg. 2015. "Understanding psychological reactance." *Zeitschrift für Psychologie* .

Thaler, Richard H and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Tomz, Michael. 2007. "Domestic audience costs in international relations: An experimental ap-

proach." *International Organization* 61(4):821–840.

Tomz, Michael and Jessica L Weeks. 2018. "Human Rights and Public Support for War.". **URL:** *http://kingcenter.stanford.edu/sites/default/files/publications/WP1026.pdf*

Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American political science review* 107(4):849–865.

Valentino, Nicholas A, Ted Brader, Eric W Groenendyk, Krysha Gregorowicz and Vincent L Hutchings. 2011. "Election nights alright for fighting: The role of emotions in political participation." *The Journal of Politics* 73(1):156–170.

Vrij, Aldert, Christian A Meissner, Ronald P Fisher, Saul M Kassin, Charles A Morgan III and Steven M Kleinman. 2017. "Psychological perspectives on interrogation." *Perspectives on Psychological Science* 12(6):927–955.

Weisiger, Alex. 2013. *Logics of War: Explanations for Limited and Unlimited Conflicts*. Cornell University Press.

Willard-Foster, Melissa. 2018. *Toppling Foreign Governments: The Logic of Regime Change*. University of Pennsylvania Press.

8 Appendix

Table 2: Regression results

	Turn Back (Binary)		Support War		Support Sanctions	
	Logit		OLS		OLS	
	(1)	(2)	(3)	(4)	(5)	(6)
Natural Consequences	−3.400**	−3.737**	−0.052	−0.053*	−0.155**	−0.150**
	(0.345)	(0.372)	(0.028)	(0.024)	(0.027)	(0.027)
Coercion	−0.950**	−1.031**	0.102**	0.075**	0.062*	0.052
	(0.213)	(0.236)	(0.028)	(0.025)	(0.028)	(0.028)
Militarism		−1.852**		0.470**		0.159**
		(0.593)		(0.052)		(0.057)
Isolationism		1.107*		−0.054		0.029
		(0.474)		(0.042)		(0.046)
Party ID		1.468*		0.062		0.059
		(0.651)		(0.057)		(0.062)
Ideology		−1.005		−0.163**		−0.075
		(0.705)		(0.060)		(0.066)
Male		−0.644**		0.037		0.044*
		(0.226)		(0.020)		(0.022)
White		−0.407		−0.041		−0.010
		(0.251)		(0.023)		(0.025)
Age: 18-24		−0.523		0.086		−0.072
		(0.705)		(0.064)		(0.070)
Age: 25-34		−1.215		0.089		−0.078
		(0.634)		(0.058)		(0.063)
Age:35-44		−1.061		0.035		−0.090
		(0.644)		(0.059)		(0.064)
Age: 45-54		−0.459		0.037		−0.027
		(0.653)		(0.060)		(0.066)
Age: 55-64		−0.768		0.056		−0.043
		(0.692)		(0.064)		(0.070)
University		−0.478*		−0.008		0.020
		(0.237)		(0.021)		(0.023)
Income Quartile 1		0.087		0.021		−0.046
		(0.330)		(0.031)		(0.033)
Income Quartile 2		−0.141		0.028		−0.022
		(0.314)		(0.029)		(0.031)
Income Quartile 3		−0.301		0.038		−0.042
		(0.305)		(0.028)		(0.030)
Constant	0.470**	2.193*	0.332**	0.152	0.658**	0.643**
	(0.152)	(0.931)	(0.020)	(0.085)	(0.020)	(0.092)
N	590	590	590	590	590	590
R^2			0.053	0.291	0.110	0.153

*p < .05; **p < .01

Note: Models 1 and 2 display logit coefficients for a binary dependent variable, Models 3-6 display OLS coefficients. Reference category for treatment dummies is natural consequences. Continuous independent variables and dependent variables have been rescaled to range from 0 to 1. Higher values on party id and ideology are more Republican and conservative, respectively.