Coercion and Provocation

Allan Dafoe[1,2], Sophia Hatz[3], and Baobao Zhang[1,4]

[1]Future of Humanity Institute, University of Oxford
[2]Department of Politics and International Relations, University of Oxford
[3]Department of Peace and Conflict Research, Uppsala University
[4]Department of Government, Cornell University

July 21, 2020

Forthcoming in the *Journal of Conflict Resolution*

Abstract

Threats and force, by increasing expected costs, should reduce the target's resolve. However, they often seem to increase resolve. We label this phenomenon provocation. We review instances of apparent provocation in interstate relations and offer a theory based on the logic of reputation and honor. We also consider alternative explanations: confounding or misimputation of resolve, revelation of information, character, or capabilities, or generalized sunk cost reasoning. Using survey experiments we systematically evaluate whether provocation exists and what may account for it. We employ design-based causal inference techniques — a hypothetical natural experiment, a placebo treatment, and ruling out mediators — to evaluate our key hypotheses. We find strong evidence of provocation, and suggestive evidence that it arises from considerations of honor, vengeance, and reputation. Our experimental design minimizes the risk that this result arises from our alternative explanations.

The authors can be reached at allan.dafoe@politics.ox.ax.uk, sophia.hatz@pcr.uu.se, and baobaozhangresearch@gmail.com. The replication data and code are found at `https://doi.org/10.7910/DVN/0BFI9L`. For helpful input we thank Tyler Bowen, Devin Caughey, Matthew Cebul, Kristine Eck, Guy Grossman, Håvard Hegre, Susan Hyde, Simon Hug, Josh Kertzer, Deborah Larson, Jason Lyall, Ed Mansfield, Zeev Maoz, Rose McDermott, Nuno Monteiro, Magnus Öberg, Jon Pevehouse, Brian Rathbun, Jonathan Renshon, Milan Svolik, Gerald Schneider, Nina von Uexküll, Steve Weber, Jessica Weiss, Remco Zwetsloot, the attendees of the 4th East Asian Peace conference, the UCLA conference on Preferences for Conflict, ISA 2015, ENCORE 2015, the Browne Center Speaker Series at the University of Pennsylvania, and especially Ken Schultz, Stein Tønnesson, Robert Trager, and Jessica Weiss.

1 Overview

A fundamental assumption in our understanding of conflict is that threats and force function as a means of coercion. The logic is simple: the use of threatened or actual harm increases a target's actual and expected costs of resistance, reducing the target's resolve, or determination to resist. And yet, contrary to this straightforward logic, threats and violence often seem to result in provocation; they appear to increase the resolve of a target, in some cases leading to greater resistance and retaliation. The September 11 attacks were arguably an attempt by al-Qaeda to persuade the US to reduce its military presence in Saudi Arabia and support for Israel. And yet, far from coercing the American public against involvement in the Middle East, the attacks of 9/11 provoked the US into invading *two* countries in the Middle East, as well as maintaining its previous commitments. Historians and scholars have documented many such examples throughout history and across cultures, in interstate and intrastate conflicts.[1] In many tense areas of the world — such as the South and East China Seas, the Balkans, Korea, and India/Pakistan — analysts express concern that even a minor "provocation," such as the collision of military planes or an exchange of fire across borders, could spiral out of control into a war.

Given our assumptions about the coercive power of actual and threatened harm, how can we explain its apparently ubiquitous provocation effect? This paper addresses the puzzle of coercion and provocation. While the phenomenon that threats and violence can provoke resistance and retaliation is recognized [2], and theories describe why actors might use threats and violence to provoke [3], we lack an explanation for provocation as a reaction in a target. GStudying the phenomenon of provocation is also complicated by observational challenges: increased resolve is unobservable, inferring resolve from actions is problematic, and historical counterfactuals are elusive. In this paper, we offer a theory for how and why provocation occurs based on a review of historical ex-

[1] O'Neill 1999; Carlisle 2007; Kalyvas 2006; Kagan 1995.

[2] Lichbach 1987; Myerson 2007.

[3] Kydd and Walter 2006; de Mesquita and Dickson 2007; Rocco and Ballo 2008; Carter 2016.

amples, inductive theory-building, and scenario-based survey experiments. We employ principles of design-based causal inference in the design and analysis of our experiments in order to evaluate the processes by which a minor "provocation" (the aggressive maneuvering of military planes) translates into public support for escalation.

We first review salient historical examples of *apparent provocation*: cases where the application of harm in a coercive setting — such as a threat, aggressive maneuvering, or an attack — appeared to substantially increase the target's *resolve* or desire for retaliation. From this review we identify an explanation for provocation that we focus on in this paper: concern for reputation and honor (E_1). E_1 emerges from the observation that threats, aggression, and violence are often perceived as tests of resolve and matters of national honor. Perceived in this way, coercive acts often put reputation (for resolve) and honor at stake. By resisting coercion, targets signal to others that they will not allow themselves to be coerced and that coercion is not effective against them. This reputational logic could emerge from many different processes. It could emerge from rational calculation of the value of a deterrent reputation[4]; from cultures that socialize and institutionalize concern for honor and revenge[5]; from psychological-emotional processes related to negative reciprocity, fairness[6], humiliation[7], and outrage.[8] We also identify three alternative explanations: confounding or the mis-imputation of resolve (E_2), information revelation (E_3), or generalized sunk-cost reasoning (E_4).

The four types of explanations we identify can be illustrated with the example of the attack on Pearl Harbor. The Japanese attack is widely perceived to have substantially increased the US public's willingness to wage war against Japan and the Axis.[9] (E_1): Perhaps the attack on Pearl Har-

[4]Weisiger and Yarhi-Milo 2015, §10; Sartori 2002.

[5]Stein 2015; O'Neill 1999.

[6]Gottfried and Trager 2016.

[7]Barnhart 2016.

[8]Hall 2016.

[9]Larson 1996, 14–15.

bor put American reputation and honor at stake, thereby activating an impulse (rational, cultural, psychological, emotional) to resist and punish Japan to restore American reputation and honor? Alternatively, it could be that the attacks only appear to have provoked, due to (E_{2a}) confounding or (E_{2b}) mis-imputation of resolve. Confounding could occur if we compare historical cases in which a provocative act occurs against cases in which one did not; these cases will differ in other ways, such as the presence of a conflict of interest or the value of the stakes. E_{2a}: Perhaps the public's perception of the stakes of the global conflict had increased over this time, coinciding with, but not caused by, the attack on Pearl Harbor? Mis-imputation of resolve could occur because in history we mostly observe actions (using force or not), we don't observe willingness to carry out actions (willingness to use force: resolve). E_{2b}: Perhaps the US public was always resolved — willing to go to war with Japan to defend the status quo — but until then it had not been necessary because the status quo had not been explicitly challenged? A second alternative explanation is that coercion could reveal information (E_3) about the coercer (the agent carrying out the coercive act)—such as about interests, character, or capabilities — which increase the target's resolve. E_3: Perhaps the attacks revealed that Japan had unacceptably expansionist ambitions in Asia, had fundamental conflicts of interest with the US, was of a low moral character, or was unacceptably powerful? A final alternative explanation is that coercion and harm incur sunk costs, engaging a tendency in individuals to increase their investment in a course of action (resolve) in an attempt to recuperate prior losses (E_4: generalized sunk cost reasoning). E_4: It could be that casualties suffered at Pearl Harbor induced a general sunk-cost framing in which winning the conflict was the only way to recuperate those past losses.

To evaluate these explanations, we employ a set of scenario-based survey experiments fielded in the United States. Our strategy for inference involves designing specific scenario-based survey experiments that provide distinct testable implications of E_1, while controlling or testing for the other explanations. To an online sample of U.S. citizens, we presented scenarios describing a realistic confrontation between U.S. and Chinese military planes in the East China Sea, in which a U.S. fatality occurs (or does not occur), and we measured the respondents' *preferences* about

how their state should behave. Thus, we (largely) overcome the problem of confounding (E_{2a}) by constructing precise scenarios that hold other features of the dispute constant, and we overcome E_{2b} by directly measuring resolve. Fundamentally, we find that provocation exists: reading about an attack on the US military increased respondents' level of resolve across multiple measures. Respondents exposed to the *attack* scenario preferred higher levels of escalation compared with the control group ($p < 0.001$). The attack scenario increased the percent of respondents who were willing to use military force to back up the United States' claims, to 64% from a control baseline of 40%. Furthermore, the attack scenario elevated respondents' tolerance for economic loss, military causalities, and risk of war.

In order to control for information revelation (E_3), so that we can evaluate the logic of reputation and honor (E_1), we embedded a *hypothetical natural experiment* in our survey scenario.[10] Specifically, we described a context in which a coercive fatality occurred as-if randomly: opposing military planes were flying dangerously close to each other and then almost collided (control) or collided, killing the US pilot (treatment). By making the two experimental conditions — *no collision* vs. *collision* — as similar as possible, except for the occurrence of a fatality in a coercive context, we reduced the informativeness of our treatment about background characteristics such as military capabilities or military intent. Our results validate this design: respondents' beliefs about military capabilities and military intent were balanced across the *collision* and *no collision* conditions, whereas they were imbalanced across the *attack* and *control* conditions. Consistent with E_1, we still observe an increase in respondents' resolve after they read about the as-if random collision. The treatment caused respondents to prefer higher levels of escalation ($p = 0.002$) and increased their tolerance for economic costs ($p = 0.044$).

To further isolate concern for reputation and honor as a mechanism, we distinguish this form of reasoning from generalized sunk cost reasoning (E_4). While sunk cost reasoning stipulates increased resolve following any kind of sunk cost, the logic of reputation and honor makes a similar prediction for the experience of *coercive* costs (harm). The experience of harm in a coercive con-

[10] On this technique, see Dafoe, Zhang, and Caughey 2017.

text, such as from the death of soldiers, engages reputation and honor, which increases the costs of later backing down. It follows that to distinguish between concern for reputation and honor and generalized sunk cost reasoning (E_4), we must evaluate whether the provocation that we are observing could be attributable to the experience of any kind of costs, or whether provocation only arises from coercive costs which put reputation and honor at stake. To this end, we constructed a *placebo treatment*, in which a US military fatality occurred during deployment in the East China Sea, but not within a coercive context. Respondents were told that a US pilot died when his plane crashed in a storm during a routine patrol in the East China Sea. The fatality was in pursuit of the same general foreign policy goals but was not due to a direct threat, aggressive behavior, or use of force. Consistent with E_1, and contrary to E_4, we find that respondents did not become more resolved, according to any of our four measures, after being told about this non-coercive fatality (as compared with the control group).

Finally, we attempt to *rule out potential mediators*, a form of mediation analysis which stays close to the experimental design. We acquire evidence about what particular considerations were important for the respondents' support for standing firm in this dispute. "Avoiding the risk of conflict that could cost the lives of Americans, Chinese and others" was the dominating concern, followed by "preserving US access to the East China Sea," "resisting China's threat to US national interests and/or national security," "maintaining the US's reputation," and "honoring the US pilot who died so that his death will not have been in vain," and more distantly "punishing China." Among these, our coercive treatments *attack* and *collision* had a large effect on the importance of "punishing China" and "honoring the US pilot," a suggestive effect on reputational concerns, and reduced the importance of the risks of war. This provides evidence that honor and reputation (E_1) are important for provocation, though more work remains to parse out the specific channels.

To summarize, we first note the apparent ubiquity of provocation in history. We confirm, using our survey experiments, that provocation does, in fact, exist: coercive threats and force can increase the resolve of the target, and this cannot solely be attributed to the revelation of information, confounding, or generalized sunk cost reasoning. The logic of reputation and honor provides a

persuasive account of our results.

The existence of provocation has important implications for our understanding of international relations.[11] First, the size of the effects that we observe are sufficiently large that it is plausible that this kind of provocation could account for the large swings in public support for war that follow highly coercive acts, such as the surprise attacks at Pearl Harbor and on September 11. Second, our findings lend support to the concern that a collision or other relatively minor incident in a tense region of the world, such as in the East or South China Seas, could provoke a war. Third, provocation is plausibly a source of domestic audience costs, and potentially the most substantively important source: leaders will pay domestic costs for backing down after experiencing a provocative event. This supports a reputation- or honor-based explanation for domestic audience costs, consistent with Fearon's original conceptualization.[12] Fourth, rally effects may themselves be better understood as also following the logic of reputation and honor: individuals will rally together and behind their leader after an event that puts the groups' reputation and honor at stake, but the support for the leader is conditional on the leader defending the groups' reputation and honor. Fifth, provocation is plausibly the basis for, or co-created with, the international and domestic norms about legitimate justifications for launching a war — *casus belli* — which often involve pointing to provocations of the opponent. Sixth, a better understanding of provocation is necessary for us to understand coercion: when coercion is most likely to work and why it takes the forms that it does. For example, it could make sense of why coercive demands are often first expressed privately[13], and when made publicly are so rarely made explicitly[14] but instead are implied through indirect diplomacy and symbolic idioms.[15]

[11] Cho 2016, Hall 2016, Weiss and Dafoe 2016.

[12] Fearon 1994, 581.

[13] Kurizaki 2007.

[14] Sechser 2011.

[15] O'Neill 1999.

2 Theory

2.1 Coercion and Cost Sensitivity

Coercion involves the use of threats, aggression, violence, material costs, or other types of threatened or actual harm, as a means of influencing a target's behavior.[16] The logic of coercion stipulates that the coercive efficacy of a given tactic — the efficacy of a threat or use of force as a means of influencing a target's preferences — depends on the extent to which the action credibly manipulates the target's cost-benefit calculus, and in particular the target's utility for resistance. When a tactic credibly increases the anticipated costs of resistance beyond the anticipated benefits, acquiescence with the coercer's demands should become preferable to resistance. Essential to the logic of coercion, then, is an assumption of *cost sensitivity*: all else equal, an increase in the costs of resistance will reduce the target's preference for resistance (the target's resolve). The assumption of cost sensitivity is embedded in many theories of war such as the balance of power theory[17], the bargaining model of war[18], and some explanations for the democratic peace.[19] The logic of coercion, and, in particular, the assumption of cost sensitivity, naturally leads to the conclusion that threats and force are effective as a means of coercion. To the extent that threats and force increase the anticipated costs of resistance[20], the application of threats and force should cause cost-sensitive individuals to

[16]The concept of coercion is complex. Appendix D offers supplementary theoretical discussion, including of the pros and cons to including "brute force" in the conceptualization of coercion (Anderson 2017), the precedent for doing so, and the subtle distinction with other activities like offers of positive inducements and information (Oye 1993, Chapter 3).

[17]Slantchev 2011, 177.

[18]Powell 1999, 111; Fearon 1995, 388.

[19]Valentino, Huth, and Croco 2010, 529; Byman and Waxman 2002; Russett 1990, 46.

[20]To be theoretically precise, what matters for coercion against a rational agent is a change in the expected costs of resistance. However, due to failures of rationality or credible communication of a threat, in practice actual harm is often needed to induce compliance. Thus, in practice experience

become less resolved.[21]

The logic of coercion and the cost sensitivity assumption is essential to theories describing public opinion and war.[22] Decades of public opinion research finds that the American public responds to national-level threats and harm in a cost sensitive way. This is the bottom line of "casualty sensitivity theory" — a theory generally considered to represent the scholarly consensus on the determinants of public support for war.[23] Casualty sensitivity theory describes public willingness to use military force as the result of a cost-benefit calculation: a weighing of the war's costs (measured primarily in casualties or fatalities) against the benefits (the stakes of the dispute and likelihood of success.[24] Scholars have noted that support for a mission seems to decline with the logarithm of casualties: just a few initial casualties can have a large negative impact on support, but as the number of casualties accumulates, the marginal effect of each additional casualty on reducing support for war becomes smaller.[25] The observation that even a few fatalities can turn US public opinion against a mission (the "Somalia Syndrome") has led some to characterize the American public as "casualty averse" or "casualty phobic." The focal question facing this literature is not *whether* casualties will reduce support for a given mission, but "under what conditions [they will do so] rapidly or more slowly?"[26]

In summary, the logic of coercion assumes cost sensitivity, and the prevailing theories of public opinion and war emphasize the cost sensitivity of publics to even low numbers of casualties (coercive harm). Absent any other countervailing effects of coercive harm, so that the predictions of cost sensitivity are *dominant*, our understanding of current theories of coercion, public opinion and war leads to a simple testable hypothesis:

---

of harm is often a good proxy for an increase in the target's perception of future harm.

[21] For examples of this assumption across theories, see Appendix A.

[22] Larson 1996; Gelpi, Feaver, and Reifler 2009*a*.

[23] Gelpi, Feaver, and Reifler 2009*a*, 8.

[24] Larson 1996; Gelpi, Feaver, and Reifler 2009*a*.

[25] Mueller 1973; Larson 1996; Gartner and Segura 1998; Gartner 2008.

[26] Gelpi, Feaver, and Reifler 2009*a*, 11.

H_{CS1} (Cost Sensitivity, Dominant) : *An increase in the expected costs of conflict, such as through being the target of a coercive threat or harm, decreases target resolve.*

2.2 Apparent Provocation

In contrast to the prominent prediction that an increase in the expected costs of resistance will **decrease** resolve, there are many examples throughout history where a threat, an attack, or other aggressive action, sometimes inadvertent, has apparently led to a dramatic **increase** in public resolve. We call this phenomenon *apparent provocation*: the seeming tendency for certain coercive acts to increase resolve and desire for retaliation. Before theorizing provocation, we first review examples and evidence of apparent provocation. We emphasize U.S. examples, but abundant evidence can be found in cases involving other countries.

Many wars in US history are generally considered to have been precipitated by an attack on the United States. The *Chesapeake-Leopold Affair*, a British attack on the *USS Chesapeake* that left three American sailors dead, is considered by scholars and historians to have sparked the War of 1812.[27] It was perceived by the American public "as a national insult, an affront to sovereignty."[28]

Propelled by the American people's rallying-cry "Remember the Maine, to Hell with Spain!" America declared war on Spain a few months after the sinking of the *USS Maine*. Some scholars note that it was the German attacks on the *Vigilancia* and *Lusitania* which provided a cause sufficient to mobilize American support for the Great War against Germany.[29] America's entry into World War II came only after the attack on Pearl Harbor (1941) shifted public opinion from opposing to favoring war against the Axis powers. Retaliation for the attack on Camp Holloway (1965), near Pleiku, sparked a major escalation in America's attacks against North Vietnam. More

[27]Ferrell, Horsmann, Perkins in Cray 2005.

[28]Cray 2005, 446.

[29]O'Neill 1999, 103; Carlisle 2007.

recently, the September 11 attacks are widely perceived as having unified the American public, the US Congress, and, to a lesser extent, the international community in common support of a military incursion in Afghanistan with the explicit mission to eliminate the perpetrators of the attacks. The September 11 attacks were also used to mobilize public support for the Iraq War (2003), a war that should have been highly unpopular due to its high expected costs and the unclear level of threat posed by Iraq.[30] In short, across an array of contexts and time periods, the use of force by an aggressor has often apparently increased the resolve of the target, provoking war.

A close reading of historical examples suggests that casualties, and especially fatalities, are the nexus of provocation. In the *Chesapeake-Leopold* Affair, several days after the attack, one of the injured American sailors died. His burial became the occasion for a large patriotic event, with thousands of people in attendance.[31] Incidents like this suggest that there is something distinctly provocative about death, above and beyond the immediate characteristics of the coercive incident. There are many examples in modern history of a small number of fatalities provoking a sudden military surge or spike in popular support for action: the Zar'it-Shtula Incident, in which three Israeli soldiers were killed, leading to the 2006 War in Lebanon[32]; the Fallujah killing of four US contractors (2004) was followed by a surge in US military operations[33]; the killing of three Israeli teenagers (2014) led to Operation Protective Edge, an Israeli invasion of Gaza. In September 2014, the Islamic State's (IS) beheading of two American journalists caused a reversal in public support for American military action in Iraq.[34]

Across different countries and cultures, the logic of provocation is often explicitly and clearly articulated. For example, in response to ISIS killings, various leaders said: "We will not be intimi-

[30]Gershkoff and Kushner 2005; Krebs and Lobasz 2007.

[31]Cray 2005, 446.

[32]Jervis 2013, 330-331.

[33]Boettcher III and Cobb 2006, 831.

[34]Dyer September 5, 2014.

dated. [ISIS's] horrific acts only unite us as a country and *stiffen our resolve*"[35]; "This will not lead Britain to shirk our responsibilities ... it must strengthen our resolve"[36]; "[the] terrorists that are capable of such atrocities need to know that they will be confronted by a France that is determined, unified and together. France will not let itself be overawed ..."[37]; "Japan will never give in ... Japan will resolutely carry out its duty as part of the international community fighting terrorism"[38]; "[The killing] will only strengthen us and reinforce our unity".[39]

Policy elites seem to perceive fatalities as capable of provoking the target to greater resolve and commitment. In 1910, when French General Foch was posed with the question of what would be the "smallest British military force of any practical assistance," he replied, "A single British soldier — and we would take good care that he was killed."[40] Thomas Schelling pointed out that the purpose of American troops in Berlin during the Cold War was to "represent the pride, the honor, and the reputation of the United States government," and, in the event of an invasion, "to die heroically ... and in a manner that guarantees that the action cannot stop there."[41] Undersecretary of State George Ball warned President Johnson that exposing US troops to combat in Vietnam would increase the US commitment to Vietnam: "The decision you face now is crucial. Once large numbers of US troops are committed to direct combat, they will begin to take heavy casualties... Once we suffer large casualties, we will have started a well-nigh irreversible process. Our involvement will be so great that we cannot — without national humiliation — stop short of achieving our complete objectives."[42]

---

[35]Pace September 3, 2014; emphasis ours.

[36]Pace September 3, 2014.

[37]Heneghan November 13, 2015.

[38]Hayashi February 1, 2015.

[39]Reed February 4, 2015.

[40]Callwell 1927, 78-79.

[41]Schelling 1966, 47.

[42]Sheehan and Kenworthy, 1971, memo dated July 1, 1965, in Staw 1981, 578.

In addition to overt attacks and killings, other kinds of coercive acts seem able to provoke. Provocative aggression includes dangerous activity such as maneuvers by military vessels or planes that could cause a collision, close surveillance ("buzzing"), simulated attacks, "accidental" firing during military exercises, transgressions into military territory or airspace, and other forms of harassment.[43] Leadership responses and media coverage of these incidents suggest that they are interpreted as challenges: as tests of the target nation's capability, readiness, and willingness to use military force in a dispute. The intrusion of Russian military planes in Swedish airspace in September 2014 was described by the Swedish Defense Force as a "test of Swedish readiness."[44] Reflecting on Chinese actions in the East and South China seas, analysts often express interpretations such as: "China appears to be testing the resolve of the Americans to defend US interests, and those of its allies, in the region."[45]

Coercion can be provocative even if it does not inflict any immediate harm. In particular, coercive demands — *threats* — often provoke. For example, the XYZ Affair, which involved a French demand for bribes from American diplomats, "set off a near hysterical reaction in the United States"[46], leading to anger at the French and support for a strong military, exemplified by the slogan "Millions for defense but not one cent for tribute." The provocativeness of threats plausibly accounts for why explicit threats are so rare in international relations, relative to mutually understood coercive demands, and why threats are usually disguised and cushioned using indirect diplomatic language.

Acts of disrespect seem to sometimes provoke states into war. Many scholars note that war can be triggered by "symbolic affronts," such as the refusal to dip a flag, to salute colors, or to

[43]Lynn-Jones 1985.

[44]Dagens Nyheter, September 20, 2014.

[45]BBC Asia May 31, 2014.

[46]Herring 2008, 86.

follow protocols[47], and "symbolic challenges" such as rude gestures or explicit verbal statements.[48] England's impressment of American sailors was viewed by the United States as a "symbolic denial of [the US's] sovereign status" and contributed to the outbreak of the War of 1812.[49] Bismarck famously edited the report of a conversation between the Prussian King and French Ambassador to Prussia to make the exchange appear insulting — a "red rag to taunt the Gallic bull" according to Bismarck. He successfully provoked France into declaring the Franco-Prussian War.[50]

Events and actions can provoke even when the facts are unclear. At the time of the sinking of the *USS Maine*, it was unclear whether the sinking was an accident or caused by an attack by Spain. Nonetheless, the American public was outraged. In other cases, an increase in resolve seems to occur before the identity of the perpetrator is known or before there is strong evidence of culpability, as in the killing of the Israeli teenagers (2014) and the MH17 shootdown (2014). Because an incident can be provocative even when culpability is uncertain, leaders can strategically portray incidents in a manner to make them more provocative, or even fabricate provocative incidents. These can then serve as pretexts for escalation. The second Tonkin Gulf attack (1946), even though it never happened, was used to galvanize support for greater US involvement in Vietnam.[51]

The provocativeness of an incident ultimately depends on perceptions: "The relevant question is not whether the honor of states is challenged, but whether they see it as challenged."[52] Nevertheless, perceptions are influenced by real-world events, which is why in extreme cases, states or groups have actually orchestrated "false-flag" events: coercive acts, even fatal attacks, against oneself and allegedly committed by the adversary, in order to mobilize support for war. In 1788 the Swedish Riksdag agreed to support the declaration of the Russo-Swedish War after an appar-

[47] Pinker 2011, 232.

[48] O'Neill 1999, 101.

[49] O'Neill 1999, 101.

[50] Wawro 2005, 37.

[51] Beschloss 2001, 38-39, in Lai and Reiter 2005, 269.

[52] O'Neill 1999, 102.

ent Russian attack on the Swedish town Puumala. The attack, however, had been staged by the Swedish king.[53] Operation Northwoods, a 1962 top-secret proposal by the U.S. Joint Chiefs of Staff, outlined "several forms" of false-flag "'Remember the Maine' incident[s]' that could provide a "legitimate provocation as the basis for US military intervention in Cuba." These included blowing up ammunition, burning aircraft, and sinking a US ship in Guantanamo Bay base; developing "a Communist Cuban terror campaign" in the United States; the "use of MIG type aircraft by US pilots [to] provide additional provocation"; and the creation of "an incident which will make it appear that Communist Cuban MIGs have destroyed a USAF aircraft over international waters in an unprovoked attack."[54] The Mukden/Manchurian incident (1931), the Gleiwitz incident (1939), and the Shelling of Mainila (1939) are similar instances of "false-flag" operations.

Similarly, hawkish groups sometimes intentionally provoke their adversary into taking escalatory actions. We see this in the 1964 attacks by al-Fatah against Israel, which provoked Israeli reprisals. This had the intended effect of provoking Arab states into hostilities against Israel, eventually culminating in the Six-Day War.[55] One interpretation of Bin Laden's motives for organizing the attacks of 9/11 was to provoke the US into a costly war in the Middle East.[56] That violence can provoke retaliation, which then advances extremists' goals, is an important feature of strategic explanations for terrorism.[57]

As the above examples illustrate, acts of coercion often seem to provoke the target to a greater willingness to resist and retaliate. Inductively generalizing from our review, incidents seem to be more provocative the more they appear aggressive, damaging and especially fatal, disrespectful, explicit, public, intentional, and not apologized for. To be clear, an incident can provoke even if it

[53]Nordstrom 2000, 119.

[54]Blouin and Ingelido 1962; Hellman 2013.

[55]Oren 2003, 1, 25.

[56]Bergen 2006, 360.

[57]Pape 2003; Kydd and Walter 2006; de Mesquita and Dickson 2007; Abrahms 2008; Carter 2016.

is not intentional–such as deaths from an unauthorized or mis-implemented military action–though usually an apology and costly compensation would reduce the provocation.

These observations lead to our primary alternative hypothesis:

H_{P1} (Provocation): *Coercive incidents can increase target resolve.*

This hypothesis only claims the existence of provocation, that there are some kinds of coercive acts, some contexts, and some targets for which coercion will increase target resolve. Nevertheless, our claim in this paper is stronger. We argue that coercion provokes an impulse to resist and retaliate, and that this impulse often overwhelms countervailing cost considerations. While our survey experiments can only establish the existence of provocation, we theorize about the factors that make coercion more or less provocative. In the conclusion of this paper, we outline how future research could examine the conditions under which provocation occurs.

2.3 (E_1) The Logic of Reputation and Honor

We argue that most instances of provocation are well explained by the logic of reputation and honor, where by "reputation" we are primarily referring to reputation for resolve for not backing down in disputes (Dafoe, Renshon, and Huth 2014). Provocative acts of coercion tend to be perceived as "tests of resolve" and "national insults," which put the reputation and honor of the leader and nation at stake. To protect reputation and honor requires resolute, and sometimes vengeful, behavior. To use an expected utility framing: although coercion increases the target's expected costs of fighting, it also often increases the reputational and honor costs to backing down. When the increase in the perceived reputational and honor costs from backing down are sufficiently great compared to the increase in material costs from fighting, agents will become more resolved: agents will be more willing to use force to avoid backing down on the issue.

The logic of reputational concerns has been elaborated in the study of deterrence. Agents who back down in a dispute will be perceived by others as more likely to back down in other disputes,

thus inviting new and more forceful challenges. Thus, agents have reputational incentives to stand firm so as to persuade others that it is not worth fighting them. When the value of future stakes that could be contested is sufficiently large, when a reputational logic is operative (that is, when disputes are perceived as interdependent), and when backing down is likely to be perceived unambiguously by others as such, then the reputational stakes will dwarf the material stakes. In such a world, it can be rational for agents to prefer risking a devastating war, rather than back down and concede on an issue that is materially unimportant but reputationally important.

The logic of reputation and honor is inter-subjective and often involves beliefs of higher order. Reputational incentives can follow a *first-order* reasoning process: how will you think about me if I back down? But they can also involve higher-order reasoning processes—how will you think that I will think that … you will think about me if I back down—making them all the more elusive. Reputation is often formalized as a game involving inferences about private information, but it can also be formalized as games involving multiple equilibria. In settings with multiple equilibria, "extra-game" factors such as history, culture, and psychology can play a big role in shaping behavior. Thus, even when theorized using a rationalist game-theoretic framework, reputation can be inherently cultural, historical, and psychological.[58]

This leads to the second way of understanding this phenomenon, as less a matter of instrumental calculation, and more of a product of culture, psychology, and emotions.[59] We use the term *honor* to refer to these more cultural and psychological processes. Honor means many things in many different cultures, but most have the commonality that an honorable man[60] has a "right to respect"[61] and should be willing to use violence to defend his honor, such as after being insulted or coerced.[62] As psychologist Michael McCullough notes "a psychological common denominator

[58]Dafoe, Renshon, and Huth 2014.

[59]O'Neill 1999; Lebow 2008.

[60]Honor is a highly gendered concept.

[61]Stewart 1994.

[62]O'Neill 1999.

among the severe, intentional harms that elicit the desire for revenge is that they violate the victim's sense of honor."[63] A common principle for avoiding dishonor is costly resistance and retaliation. Costly retaliation against unfair or hostile actions—referred to as the revenge instinct or negative reciprocity—is an almost universal behavioral impulse.[64] Daly and Wilson find that the specific principle of retribution of a "life for a life" is held by 57 of 60 tribes that they investigated.[65] And while the impulse for retribution appears universal, some cultures, specifically cultures that are concerned about honor, elevate the impulse for retribution to a sacred obligation.[66] The different logics of reputation and honor—rational, cultural, psychological—could express themselves at the mass-public level, at the elite level, or both. And each level could shape the other: elite institutions can shape how the public understands international events; public opinion can constrain elite behavior and shape the reputational and honor implications of events. Most historical examples and our experiments focus on public opinion, but the theory also applies to elites.

We use the term *reputation* to refer to the more instrumental strategic considerations that could generate this phenomenon, and the term *honor* to refer to the more intrinsic, psychological, emotional, and cultural processes that could generate this phenomenon. In practice, though, these different kinds of explanations are not easily disentangled. The rational pursuit of reputation leads agents to seek commitment devices, which can look like intrinsic, psychological, and cultural obsession with honor. For example, leaders can rationally act "crazy," as Nixon tried to do with his "madman theory" to coerce the Soviets.[67] Leaders can delegate responsibility to actors with more hawkish preferences or set up organizational processes that automate retaliation. Groups can adopt and promote cultures with an honor orientation, as exemplified by the US norm against negotiating with terrorists, and the discourse in the US public that valorizes US "credibility."

[63] McCullough 2008, 69.

[64] Henrich et al. 2006; Rathbun and Kertzer 2015.

[65] Daly and Wilson 1988.

[66] Stewart 1994; Nisbett and Cohen 1996; Miller 2005, 2009.

[67] Sagan and Suri 2003.

Similarly, many intrinsic, psychological, and cultural processes plausibly arose because of the instrumental benefits they conveyed over the course of the evolution of humans and cultures. Further, these *honor* processes often appear to allow for conscious strategic considerations. For example, William Miller recounts how the Icelandic Vikings, one of the world's most honor-obsessed people, would nevertheless "in the heat of battle [keep] a running account of the damage they were inflicting, deducting the damage they were suffering," lest they would have to pay too much in compensation after the battle.[68] A close friend and a biographer commented about US President Andrew Jackson — known for his quick temper, anger, and vengefulness — that he knew "when to get into a passion and when not" and that his anger was "fierce, but never had any ill effect upon his purposes; on the contrary, he made it serve him ..."[69] Examples abound of honor-obsessed individuals ignoring or laughing off an insult to which it would be especially imprudent to take offense. We thus argue here that adjudicating between the instrumental, intrinsic, and cultural explanations is fraught with complications, and is probably infeasible at this stage of research. Rather than offer a specific explanation based on some particular interaction of instrumental, intrinsic, and cultural processes, for which theory and empirics are unable as yet to adequately support, our theory highlights the logic common to these overlapping and complementary explanations, which we refer to as the *logic of reputation and honor*.

Instrumental, cultural, and intrinsic approaches to theorizing reputation and honor lead to many similar predictions about provocation. The primary implication is that the more that reputation or honor is at stake, and the more an individual cares about their (or their group's) reputation or honor, the stronger the impulse to stand firm or seek retribution. Reputation and honor are put at stake when one is the target of coercion. In particular, coercive acts put reputation for resolve (the reputation for not backing down) at stake. The meaning of being the "target of coercion" depends somewhat on culture and context, but it generally includes being the target of a threat, an attack, a competition in risk taking, or experiencing harm due to the aggressive actions of another.

[68] Miller 2005, 119.

[69] Fischer 1989, 769.

There will typically be more reputation and honor at stake when the coercive act is more explicit, public, aggressive, insulting, intentional, and not apologized for. The reputational stakes will also be greater when the reputation is of greater benefit, such as when disputes are perceived to be interdependent, the agent could face many potential challenges, and when wealth is more easily taken by force such as is the case for states with valuable natural resources.

That leaders weigh the reputation and honor costs of withdrawal is apparent in their justifications for resisting coercion. Woodrow Wilson offered a reputational argument for retaliating against German attacks preceding World War I: once we "accept a single abatement of right and many other humiliations will certainly follow."[70] Margaret Thatcher offered a reputational argument for retaliating against Argentina: "The eyes of the world are now focused on the Falkland Islands ... Wherever naked aggression occurs it must be overcome. The cost now, however high, must be set against the cost we would one day have to pay if this principle went by default."[71] Analysts of the East and South China Sea disputes have noted the role of concerns about reputation and honor among American and Chinese leadership today.[72]

To summarize, reputation and honor can be harmed by backing down after being the target of coercion, including from experiencing harm in a coercive context. This leads to the following testable hypothesis:

H_{R1} (Reputation): *Coercive incidents can increase target resolve.*

This prediction is the same as that for provocation (H_{P1}), it is in this sense that the logic of reputation and honor offers an explanation for the apparent empirical phenomenon of provocation. We now turn to some alternative explanations for apparent provocation.

[70] Tucker 2007, 201.

[71] Thatcher 1982.

[72] See, for example,Weiss 2013 and Stephens 2014.

2.4 (E_2) Confounding or Mis-imputation of Resolve

The above sections provided historical anecdotes suggesting that provocation exists, and offered an explanation for provocation based on the logic of reputation and honor. This section considers two alternative explanations that posit that provocation is an illusion: confounding and the mis-imputation of resolve.

Much of the historical evidence consisted of examples of conflicts that seemed to erupt following an initial provocation. However, it could simply be that there are unobserved causes of apparent provocations and conflict that explain their co-occurrence. So long as parties take turns escalating a conflict, and hostile coercive actions are easily seen as provocations, any factor that makes conflict more likely will generate a pattern of apparent provocation: as a conflict escalates some coercive action will be taken that can be interpreted as a provocation of the later escalation. For example, in the Pearl Harbor example, the US and Japan already had a deep conflict of interest. As this conflict escalated, some coercive action would have taken place that looked like a provocation, but it would be incorrect (under this interpretation) to say that the action caused the later escalation. The escalation of the conflict was being driven by the background conflict of interest. This *confounding* explanation is most concerning for cross-case comparisons. Suppose a scholar had data on the occurrence of provocations (the independent variable), and the occurrence of conflict escalation (the dependent variable). It would be problematic to look to their association to infer something about the effect of provocations, since provocations and escalation are both consequences of the same underlying process.

Our historical evidence went beyond cross-case comparisons, though. Most of our examples involved an implicit interrupted time-series design: the provocative event was sudden and sometimes plausibly exogenous, and the change in the outcome (resolve of the target) was large and also sudden. In order for confounding to account for these results it would have to be that the unobserved cause of conflict suddenly increased, causing both the apparent provocation and sudden change in resolve. For Pearl Harbor, for example, this is not plausible. Public support for war against Japan jumped dramatically after Pearl Harbor; there is no plausible unobserved factor that could account

for this.

There is another explanation that could account for the apparent provocation at Pearl Harbor and other interrupted time-series results. This can be called the mis-imputation of resolve, a change in the status-quo, or the activation of a latent dispute. Suppose that the US public in 1941 always preferred a war with Japan over allowing Japan to claim hegemony in East Asia and Southeast Asia. Before the attack on Pearl Harbor, the American public did not support war against Japan because the status quo appeared secure. However, after the attack, the status quo was no longer secure, and thus war became necessary to restore it. Polls would show a change in support for war, but what changed was not resolve but the status quo implicit to this question.

We call this mis-imputation of resolve because in history we typically do not observe resolve (willingness to use force over some issue), but actions (actual uses of force). Accordingly, when we see a country use force after being the target of coercion it may be that the country's resolve did not increase, but that the conflict escalated to the point where it became appropriate for the country to use force to defend its interests.

As will become apparent, our scenario-based survey experiments allow us to overcome, in this empirical setting, these two potential alternative explanations of provocation, so that we can have a more controlled test of the Logic of Reputation and Honor (E_1).

2.5 (E_3) Information Revelation: Interest, Character, and Capabilities

In addition to the logic of reputation and honor, we offer another explanation for provocation. Provocative events could reveal information about the coercer, which makes the target more resolved. An act of coercion could reveal something about interest, intent, character, or capabilities, that make it rational for the target to be more resolved or aggressive. For example, Trager points out how something as seemingly benign as a state privately informing another state (the target) about a conflict of interest could persuade the target that "it cannot achieve its key strategic aims through its current relationship," which could lead the target to "reorient [its] security policies in order to drain resources from the threatening state," to "form new alliances," and, if conflict now

seems more likely, to "increase arms production, mobilize troops, or strike first."[73] Similarly, Hall (2011) theorizes that when states project an "image of anger," through discourse, symbolic gestures or actions, this can reveal information about the emotional salience of the issue at stake and the extent to which conciliation can diffuse escalation. The "diplomacy of anger," like coercion, can influence a target's choice between conciliation and continued resistance. These arguments apply not just to explicit communication and image-projection, but to any revelation of information about a conflict of interest, such as from an insulting diplomatic exchange or military actions that violate the expectations of the relationship. Coercion very often reveals information about the existence and depth of a conflict interest; accordingly, provocation plausibly arises from the target's rational reorientation of its foreign policy given new information about this conflict of interest.

In addition to revealing a conflict of interest, a coercive action could reveal that the coercer has an undesirable character. Coercive actions are often described by leaders and in the media using language that suggests the coercer is immoral. For example, the January 2015 terror attack in France, in which 12 French citizens were killed, was described by French President Hollande as an act of "exceptional barbarity"[74] and by US President Obama as "cowardly" and "evil."[75] It seems that a coercive act is considered an "atrocity" more because of the nature of the attack and attacker than because of the extent of physical harm incurred. Similar language was used to describe the Pearl Harbor and September 11 attacks. In his Infamy Speech, the day after Pearl Harbor, President Franklin Delano Roosevelt cited the attack as evidence that Japan had been deceiving the US by participating in negotiations, all the while planning for war. *TIME* Magazine described the attacks as "premeditated murder masked by a toothy smile."[76]

Coercion could also reveal information about the capabilities of the coercer. An act of coercion could suggest that the coercer is more confident in its capabilities, or could directly demonstrate

[73]Trager 2010, 350.

[74]Viscusi and Campbell January 8, 2015.

[75]Pickler January 7, 2015.

[76]TIME December 15, 1941.

or influence its capabilities; while such information could deter, it could also provoke if the target now perceives the coercer as a greater threat to the target's interests. For example, the attacks on September 11 2001 revealed that al-Qaeda and other terrorist groups were capable of inflicting thousands of fatalities on the US homeland; al-Qaeda was no longer a problem facing only Middle East allies, but also one posing a credible threat to the lives of civilians in the United States. In addition to the revelation of information about the interests and character of al-Qaeda, the revelation of information about al-Qaeda's capabilities could be offered to explain the change in US public support for a war in Afghanistan. Alternatively, a coercive act could reveal weakness, incompetence, or desperation, all of which in some circumstances could provoke the target.

From our historical review, we believe that information revelation plausibly accounts for many instances of provocation. In this paper, we seek to specifically establish whether concern for reputation and honor (E_1) could account for (some) provocation, and thus we employ an experimental design (the embedded natural experiment) in which we are able to minimize information revelation, to isolate E_1. We study the public response to provocation because the historical cases we review suggest that provocation can increase the public's support for military conflict and thereby influence national leaders' decision to use force. In connecting mass politics with foreign policy, existing research suggests that the public could impact foreign policy by holding existing leaders accountable[77] and selecting future leaders who promise appealing foreign policy.[78]

3 Survey Experiment Design and Results

The experimental design was based on several pilot studies that generated similar results, including one in China. We then ran a confirmatory study in February 2016, the design and analysis code for which was preregistered (available at [removed for peer review]). Thus all inferential statistics

[77] De Mesquita and Siverson 1995; Gelpi, Feaver, and Reifler 2009*b*; Baum and Potter 2015.

[78] Fearon 1998; Gadarian 2010; Tomz, Weeks, and Yarhi-Milo 2017.

(p-values) can be interpreted as having their formal statistical properties, as they are demonstrably free of the risk of undisclosed specification search and multiple-comparisons. We present here the design and results of this experiment, which involved 1,761 respondents recruited from Amazon.com's Mechanical Turk. Mechanical Turk respondents are diverse, more representative than in-person convenience samples, and can produce high-quality responses.[79] Furthermore, political scientists have been able to successfully replicate survey experiments conducted on nationally representative samples using Mechanical Turk respondents.[80]

Respondents read a hypothetical scenario based on the ongoing dispute between the United States and China over the US's right to conduct military operations in the East and South China Seas. Our key experimental manipulation consists of additional information regarding a recent incident in this dispute. The *Control* group received no additional information other than the background information presented in the scenario. The *Attack* condition described China attacking a US plane, killing the pilot. The *Collision* and *Non-Collision* conditions described US and Chinese planes involved in dangerous maneuvers, and either colliding or not colliding depending on the condition. The *Accident* treatment described a Non-Coercive event (a weather accident) leading to the death of a US pilot. In addition, respondents were assigned to one of two conditions comparing American and Chinese military capability. Table 1 summarizes.

The main outcomes measured respondents' level of resolve: willingness to stand firm in the dispute. We used several measures that captured different aspects of willingness to stand firm. Respondents also answered questions about possible mediators. These mediator measures enable us to understand why provocation induces a change in respondents' resolve. For complete details see the survey text in Appendix B.2.

[79]Berinsky, Huber, and Lenz 2012; Huff and Tingley 2015; Buhrmester, Kwang, and Gosling 2011.

[80]Berinsky, Huber, and Lenz 2012; Coppock 2016.

Treatment	Treatment Level	Treatment Condition	Assignment Probability
Military Capability	Military Disadvantage (M0)	M0	50%
	Military Advantage (M1)	M1	50%
Incident	Control (I0)	M(·) + I0	20%
	Attack, 1 fatality (I1)	M(·) + I1	20%
	Collision, 1 fatality (I2)	M(·) + I2	20%
	Non-Collision, no fatalities (I3)	M(·) + I3	20%
	Accident, 1 fatality (I4)	M(·) + I4	20%

Table 1: Factorial Design of the Survey Experiment.

The background information in the scenario describes each country's claim in the disputes and each country's objectives. The United States claims the right to conduct military operations in the East China Sea and have unrestricted access to the East China Sea. China claims the East China Sea is within its Air Defense Identification Zone (ADIZ) and wants to restrict US access. The dispute is described as "tense". Respondents then read that the United States has a *military advantage* over China "for a conflict in the East China Sea," or a *military disadvantage*.

The *Control* group received no additional information. The *Attack* group reads about an overt, serious, and intentional coercive act: the adversary intentionally shoots down a military plane belonging to the respondent's country that was "trespassing," the pilot dies in the crash. Consistent with our intention that *Attack* be perceived as more coercive and costly, we find that respondents perceive the *Attack* condition as much more intentionally coercive and likely to lead to (about 10) more future American fatalities than the *Control* condition (Figure 9).

The *Collision* and *Non-Collision* treatments were cases of less explicit coercion: the adversary engages in aggressive maneuvering in a manner likely to cause harm. In the *Collision* treatment, a collision occurred due to a Chinese military plane flying dangerous maneuvers around an American plane. In the *Non-Collision* treatment, the same dangerous maneuvers *almost* resulted in a

collision. This is modeled on the 2014 incident in which a Chinese jet intercepted a US American surveillance plane; in 2001 a similar incident resulted in a collision in which the Chinese pilot died. The *Non-Collision* and *Collision* treatments mimic a natural experiment: in each case, the adversary plane made "several close passes," which almost resulted in a collision or results in a collision "on the third pass." By doing so the two conditions should reveal the same information, differing only in whether a chance event led to the experience of harm in a coercive context. As intended, respondents also perceived the *Collision* condition as likely to lead to more future American fatalities (Figure 9).[81]

Finally, the *Accident* scenario occurred within the context of the conflict, but is not itself coercive: an American military plane crashes in a storm while conducting routine exercises. Consistent with this design, respondents did not perceive any greater coercion or risk of American fatalities in this condition (Figure 9).

Before respondents proceed to the survey questions, they were given the option to open the scenario in a separate window. We recommended that respondents do so in order to be able to refer to the information in the scenario as they answer the survey.

3.1 Survey Questions

Respondents then answered several questions. The first four measured our key outcome: resolve. For our primary outcome measure, we asked respondents how far the US should go, in terms of using military force, to assert its claims in the dispute. We also asked how far the US should go in terms of: 2) future economic costs; 3) future US military fatalities; and 4) risk of war with China. Table 2 summarizes the outcome measures, while Appendix B.2 contains the complete survey text.

[81] Nevertheless, some information revelation may still be present. For example, one might infer from the collision, relative to no collision, that the adversary undertook greater risks.

Question Name	Question Text
Levels of Escalation	In terms of using military force, how far should the US go to assert its claims in this dispute? [Answer choices range from 0 to 3; 0 = The US should not assert its claims; 3 = The US should assert its claims, using military force if necessary, even if this risks war.]
Economic Cost	In terms of future economic costs, how far should the US go to assert its claims in this dispute? [Answer choices range from 0 to 9; 0 = The US should not assert its claims; 9 = The US should assert its claims no matter the economic cost.]
Military Deaths	In terms of future US military fatalities, how far should the US go to assert its claims in this dispute? [Answer choices range from 0 to 7; 0 = The US should not assert its claims; 7 = The US should assert its claims no matter the number of military fatalities.]
Risk of War	In terms of risking war with China, how far should the US go to assert its claims in this dispute? [Answer choices range from 0 to 6; 0 = The US should not assert its claims; 6 = The US should assert its claims no matter the risk of war with China.]

Table 2: Main Survey Questions: Measuring Resolve.

Next, respondents answered questions about beliefs that could mediate the effect of our treatments according to different theories. These included questions about the material costs of the dispute, the reputation-stakes of the dispute, the coercive nature of the event described, and the characteristics of the adversary (military capability and military intent). We also asked respondents about the importance of six different considerations in their decision to escalate the dispute. The order of this second set of mediator questions is also presented at random.

Lastly we asked respondents about their demographic characteristics, including their gender, political views, race, and level of education. This information is used for balance checks, data quality checks, and subgroup analysis. Summary statistics of the respondents' demographic characteristics are presented in Appendix C.1.

3.2 Existence of Provocation

We first evaluate whether provocation exists by comparing the Control and Attack conditions across the four resolve measures. As Figure 1 demonstrates (see also figures in Appendix C.4), the Attack scenario substantially increased respondents' willingness to stand firm across all four measures (detailed results in Appendix C.3.1).

In order to assert the United States' claims, respondents became more willing in the Attack scenario (relative to the Control condition) to:

- use force, from 40% to 64%
- tolerate at least a low risk of war, from 42% to 58%.
- tolerate some future economic costs, from 57% to 66%
- tolerate some military fatalities, from 27% to 38%

Increased resolve is apparent not only starting from this moderate baseline but also across the entire distribution of resolve, including at high levels. For example, the proportion willing to accept any risk of war increased from 3% to 6%, to stand firm no matter the economic cost increased from 4% to 7%, and to accept over 3000 U.S. military fatalities increased from 5% to 7%. As a basis for comparison, these estimated effects are about *six to ten times* greater than the estimated effect of military capabilities; specifically, we manipulated the description of military capabilities from "experts agree that the US's military capability is [*inferior* to *superior*] to China's for a conflict in the East China Sea." The Attack condition was perceived as much more coercive than the Control condition, and as likely to lead to many more US deaths (Figure 9).

Given the consequences of a war between the US and China, the effects described above are substantively large. A war between the US and China could be very destructive, and given how economically interdependent the two are, would probably dramatically hurt both economies. Having had a single plane shot down, Americans are more willing to endorse a policy that could lead to such a war, enough to convert a minority (40%) to a solid majority (64%) in support of using force

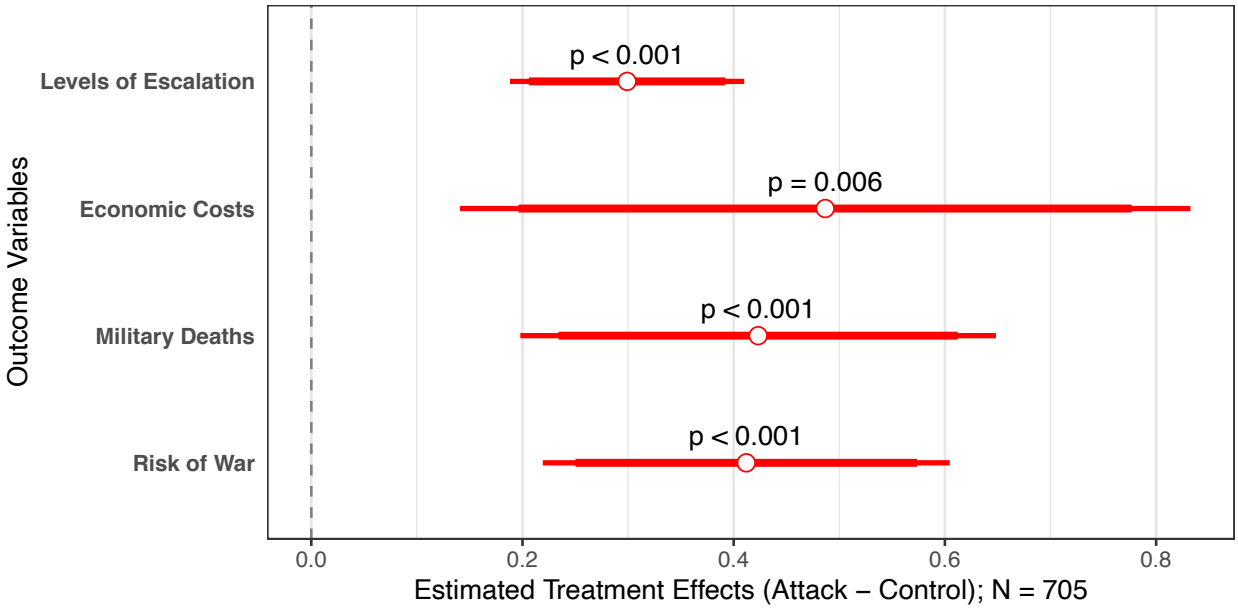

Figure 1: Attack vs. Control: Differences in Outcomes. We compare the Attack condition to the Control condition to establish the existence of provocation. Each outcome variable is a measure of respondent resolve; higher values indicate greater resolve. For each outcome variable, we show the estimated treatment effect: the difference in resolve comparing the Attack condition to the Control condition. The 95 and 99% heteroskedasticity-robust confidence intervals are shown in red.

to assert its claims. It is not hard to imagine that a more provocative event or a sequence of similar events could have even greater effects on public support for a war, generating such large swings in public support for war as was observed after the attacks on Pearl Harbor and on September 11, 2001.

3.3 (E_3) Ruling out Information Revelation

To evaluate whether provocation could be entirely due to information revelation we look for provocation under our embedded natural experiment design, the Collision vs. Non-Collision conditions. Under this scenario, the coercive fatality should arise as-if randomly. Consistent with the intended design, respondents do not appear to draw inferences about military capabilities or military intent after the as-if random collision, as seen in Figure 2. In contrast, the Attack condition induced respondents to consider China's military intent as more aggressive compared with the Control con-

dition.

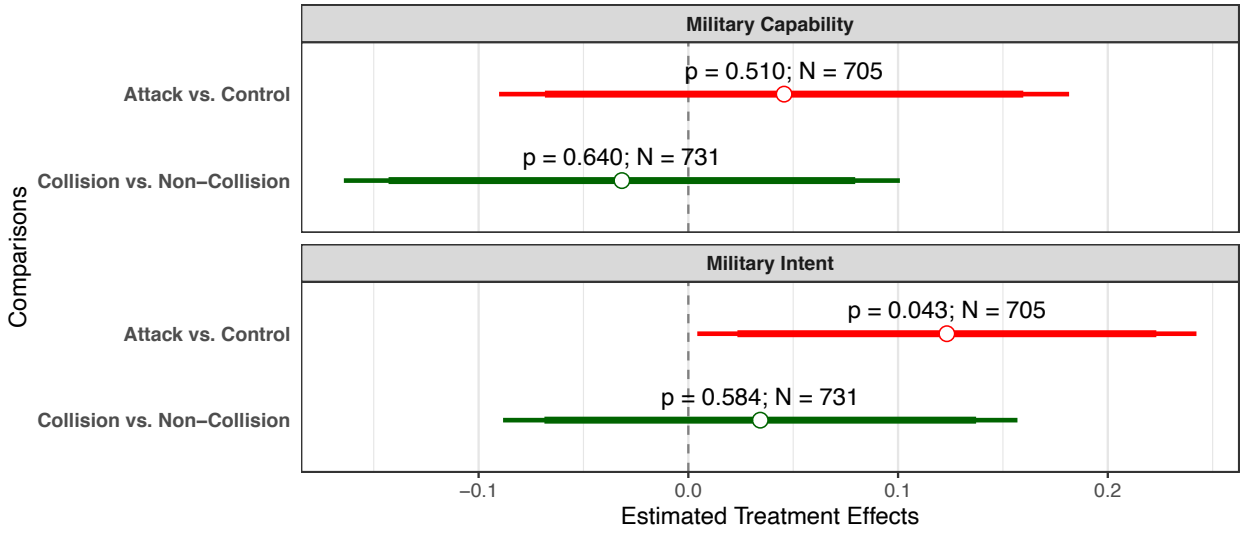

Figure 2: Placebo Outcomes. We use two placebo test questions to rule out an alternative explanation: information revelation. The outcome variables are respondents' perception of China's pre-incident military capability and military intent. Higher values indicate greater capability/intent. For each outcome variable, we show the estimated treatment effect: the difference in perceptions comparing the Attack condition to the Control condition, and comparing the Collision condition to the Non-Collision condition. The figure shows the point estimates as well as the 95 and 99% heteroskedasticity-robust confidence intervals.

Do we still see evidence of provocation when information revelation (E_3) is ruled out? Similar to the Attack vs. Control scenarios, we observe a significant increase in resolve measured by willingness to escalate and tolerance for economic cost after the as-if random collision (see Figure 3). The percent of respondents who support using military force to maintain the US's claim increased from 46% under the Non-Collision condition to 58% under the Collision condition. Likewise, the percent willing to risk economic loss increased from 58% to 66%. In addition, we observe increases in respondents' tolerance for military casualties and risk of war induced by the Collision scenario, though these effects are less significant.

3.4 (E_4) Generalized Sunk Cost Reasoning

The logic of reputation and honor implies a form of sunk cost reasoning. The experience of harm in a coercive context, such as the death of soldiers, engages reputation and honor, which increases

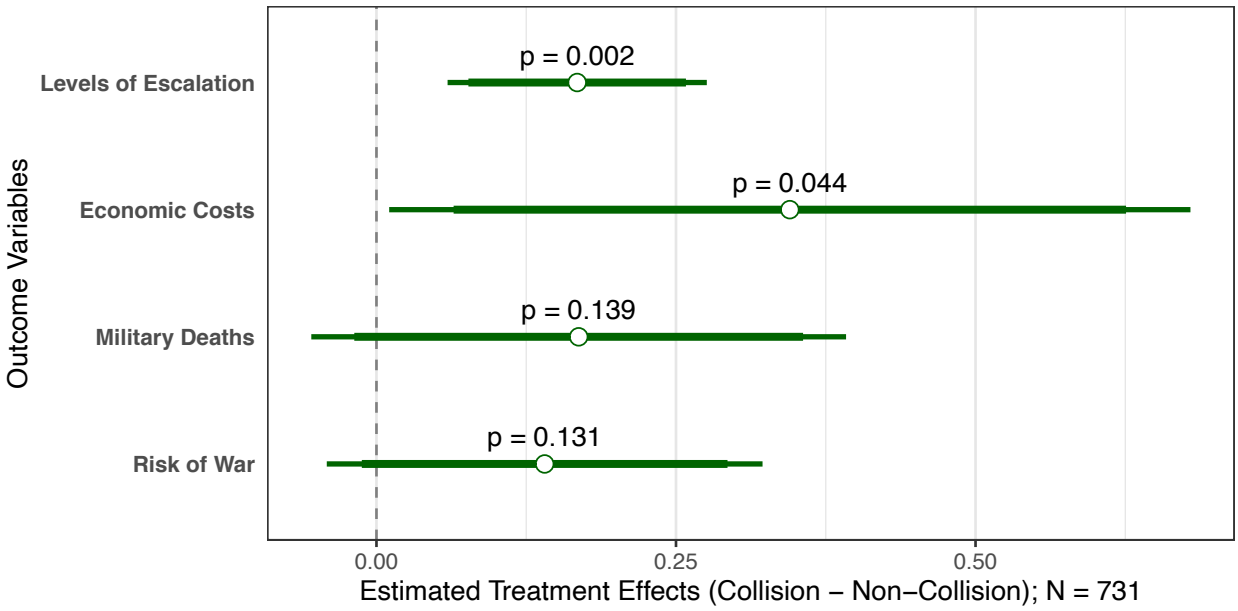

Figure 3: Collision vs. Non-Collision: Differences in Outcomes. We compare the Collision condition to the Non-Collision condition to establish whether provocation occurs in the absence of information revelation. For each outcome variable, we show the estimated treatment effect: the difference in resolve comparing the Collision condition to the Non-Collision condition. Note that the Non-Collision condition serves as the "control" condition in this comparison wile the Collision condition serves as the "treatment" condition. The 95 and 99% heteroskedasticity-robust confidence intervals are shown in green.

the costs of later backing down. Could it be, however, that provocation is driven by a generalized sunk cost reasoning, rather than by mechanisms that specifically follow the logic of reputation and honor? War casualties, in particular, could be interpreted as a sunk cost (a lost investment), because decision makers exposed their soldiers to these predictable risks, which, when contemplating walking away from the conflict without a success, could be understood as a lost investment. To evaluate this possibility we construct a scenario as similar as possible, except with a fatality that does not put reputation or honor at stake.

We argue that non-coercive fatalities are less likely to be perceived as putting reputation or honor at stake. The US policy on honoring fallen servicemen is illustrative: soldiers who die while on active duty in a designated combat zone are eligible for military funeral honors and are included in war memorials, while those who die in military accidents or commit suicide are not. For example, the names of the 74 sailors who died aboard the *USS Frank E. Evans* in an accidental collision with an Australian carrier in 1969 are not inscribed on the Vietnam Veterans Memorial (Foreign Policy, October 2, 2014).[82]

To address this concern we construct a scenario in which a pilot dies during a "routine exercise" that took place during a storm. We find that this non-coercive fatality did not increase target resolve, relative to the near collision, as shown in Figure 4. Thus, it is not the experience of losses in the pursuit of a foreign policy objective that induces this loss aversion and hardens resolve, but specifically the experience of coercive losses.

[82]Though note that reputation for cost tolerance, which is a component of reputation for resolve, could still be put at stake by the accidental death. To the extent that it does, this design will not allow us to isolate generalized sunk reasoning free of the effects of the logic of reputation and honor.

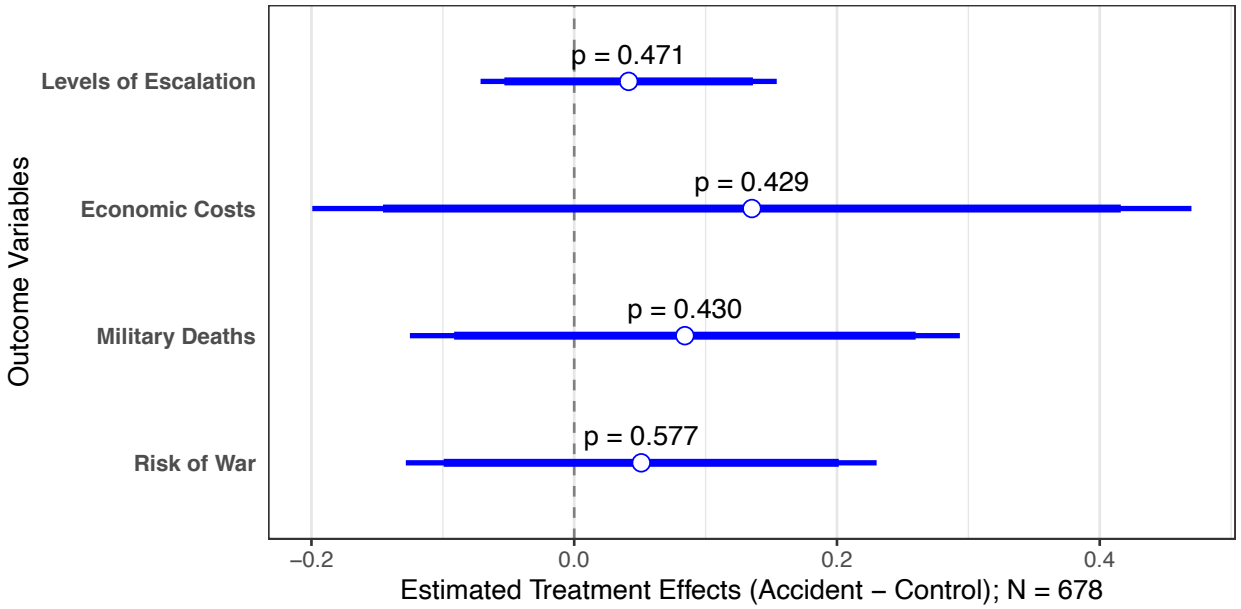

Figure 4: Accident vs. Control: Differences in Outcomes. We compare the Accident condition to the Control condition to establish whether provocation is driven by generalized sunk cost reasoning. For each outcome variable, we show the estimated treatment effect: the difference in resolve comparing the Accident condition to the Control condition. The 95 and 99% heteroskedasticity-robust confidence intervals are shown in blue.

3.5 Reasons for Provocation

The above results established that fatalities experienced in a coercive context can provoke (increase resolve), even when the fatality was inadvertent; it does not provoke when the fatality was an accident outside of the immediate coercive context; information revelation is unlikely to account for this result. As reported in Appendix C.2, we explore potential mechanisms for this effect by *ruling out potential mediators*.[83] To summarize, we find suggestive evidence of the importance of vengeance, honor, and reputational considerations in motivating provocation, facilitated by the attenuation of concern about costly conflict.

[83] Gerber and Green 2012, §10.4

4 Conclusion

Holding stakes fixed, many theories would expect that an act of coercion, which increases the real and expected costs of conflict, should reduce the target's willingness to resist.[84] Coercion works by raising the expected material costs facing the target; if the costs are sufficient the target will prefer to withdraw from the dispute rather than resist or retaliate. This view is most strongly articulated by casualty sensitivity studies, which find in decades of public opinion research that casualties decrease public support for war. Our review of historical instances of provocation and the results of our survey experiment suggest that this is not always the case. On the contrary, although acts of coercion increase the actual and expected material costs of fighting, they can also provoke an increase in target resolve.

It is not our intention to imply that the provocative effects of coercive incidents always dominate the coercive effects of these incidents. Logically it must be that coercion sometimes works, otherwise, no rational agent would ever use coercion. Further, provocation can be managed, especially if we gain a correct understanding of its causes. US President John F. Kennedy's handling of the Soviet downing of an American plane (resulting in the death of an American pilot) during the Cuban Missile Crisis effectively prevented this incident from leading to nuclear war. Kennedy realized that the event had the potential to escalate the crisis beyond his and Khrushchev's control. "They've fired the first shot, we are now in an entirely new ball game," Kennedy said.[85] Resisting pressures from American military commanders to strike back, Kennedy extended a deal that Khrushchev accepted, bringing the stand-off to a peaceful resolution. Another means of managing provocation are formalized principles for how one should react to coercive actions, namely Military Rules of Engagement (ROE); these reduce the uncertainty about how one will react to a provocative event.[86] Communication between leaders has proven effective, at times, in mitigating the provocativeness

[84] Fearon 1995; Powell 1999.

[85] Klein 2012.

[86] Hayes 1989.

of events, especially if the offending side explains and apologizes for the action. Costly and public apologies have historically reduced tensions that have arisen from provocative incidents, as in the Baltimore Crisis (1891), the Dogger Bank Incident (1904), the USS Panay Incident (1937), USS Liberty Incident (1967), the NATO bombing of the Chinese Embassy in Belgrade (1999), and the Hainan Island incident (2001). That an apology can reduce the risk of escalation is consistent with a reputation and honor based theory of provocation. An apology is often referred to as a means for the target to "save face": it mitigates the damage incurred to national honor and reduces the necessity of retaliating or escalating.[87]

Our primary aim in this paper was to show that provocation exists. Our secondary aim was to evaluate the logic of reputation and honor as a particular explanation for it. We find that the provocation in our scenarios seems to follow the logic of reputation and honor, and cannot be adequately explained away by confounding, mis-imputation of resolve, information revelation, or generalized sunk cost reasoning.

These findings set the stage for further research on the phenomenon of provocation. While we demonstrate that coercion can increase resolve and identify the logic of reputation and honor as an explanation, we do not demonstrate the conditions under which coercion is more or less provocative. The situations in which coercive harm increases resolve and those in which coercive harm compels a population to give-in remain unclear. From our review of historical cases, incidents seem more provocative when they appear aggressive, damaging and especially fatal, disrespectful, explicit, public, intentional, and not apologized for. But on the other hand, an incident can provoke even if it is not intentional and even if it causes only minor physical harm. What makes coercion more or less provocative is thus likely to depend on what the target values or associates with honor, and whether the target perceives its honor challenged. One avenue for future research could be to map out the conditions that engage reputation and honor concerns. This is likely to be culturally-specific, yet there may be certain universal triggers. This has clear policy relevance: a better understanding of the mechanisms by which coercion increases resolve will help policymak-

[87] Pruitt, Zubin, and Kim 2004.

ers manage crisis escalation.

References

Abrahms, Max. 2008. "What Terrorists Really Want Terrorist Motives and Counterterrorism Strategy." *International Security* 32(4): 78–105.

Abrahms, Max. 2013. "The Credibility Paradox: Violence as a Double-Edged Sword in International Politics." *International Studies Quarterly* 57(4): 660–671.

Achen, Christopher H, and Duncan Snidal. 1989. "Rational deterrence theory and comparative case studies." *World Politics* 41(02): 143–169.

Anderson, Scott. 2017. "Coercion." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Winter 2017 ed. Metaphysics Research Lab, Stanford University.

Barnhart, Joslyn. 2016. "The Role of the Past: Status, Humiliation and Territorial Conquest." Unpublished paper, Wesleyan University.

Baum, Matthew A, and Philip BK Potter. 2015. *War and Democratic Constraint: How the Public Influences Foreign Policy*. Princeton, NJ: Princeton University Press.

BBC Asia. May 31, 2014. "Chuck Hagel: Beijing 'destabilising' South China Sea." *BBC News* .

Bergen, Peter L. 2006. *The Osama Bin Laden I Know: an Oral History of Al Qaeda's Leader*. New York: Simon and Schuster.

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.

Blouin, F. J., and M. J. Ingelido. 1962. "U.S. Joint Chiefs of Staff Memo Operation Northwoods Justification for U.S. Military Intervention in Cuba." Unpublished paper, National Security Archive. `http://nsarchive.gwu.edu/news/20010430/`

Boettcher III, William A, and Michael D Cobb. 2006. "Echoes of Vietnam? Casualty Framing and Public Perceptions of Success and Failure in Iraq." *Journal of Conflict Resolution* 50(6): 831–854.

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling. 2011. "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" *Perspectives on psychological science* 6(1): 3–5.

Bullock, John G, Donald P Green, and Shang E Ha. 2010. "Yes, but What's the Mechanism? (Don't Expect an Easy Answer)." *Journal of Personality and Social Psychology* 98(4): 550.

Byman, Daniel, and Matthew Waxman. 2002. *The Dynamics of Coercion*. Cambridge, UK: Cambridge Univeristy Press.

Callwell, Charles Edward. 1927. *Field-Marshal Sir Henry Wilson, bart., GCB, DSO: his life and diaries*. Vol. 1 C. Scribner's Sons.

Carlisle, Rodney. 2007. "The Attacks on US Shipping that Precipitated American Entry into World War I." *Nothern Mariner* 17(3): 41.

Carter, David B. 2016. "Provocation and the Strategy of Terrorist and Guerrilla Attacks." *International Organization* 70(01): 133–173.

Cho, Hyun-Binn. 2016. "Provocation, Inadvertent War, and Military Signaling During Crises." Unpublished paper, . http://people.fas.harvard.edu/~jkertzer/HISC2017/schedule/papers/Cho.pdf

Coppock, Alexander. 2016. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." Unpublished paper, Yale University. https://alexandercoppock.files.wordpress.com/2016/02/coppock_generalizability2.pdf

Cray, Robert E. 2005. "Remembering the USS Chesapeake: The politics of maritime death and impressment." *Journal of the Early Republic* 25(3): 445–474.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2017. "Confounding in Survey Experiments." Unpublished paper, Yale University and MIT. http://www.allandafoe.com/confounding

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

Daly, Martin, and Margo Wilson. 1988. *Homicide*. New Brunswick, NJ: Transaction Publishers.

de Mesquita, Bruce Bueno, and Eric S. Dickson. 2007. "The Propaganda of the Deed: Terrorism, Counterterrorism, and Mobilization." *American Journal of Political Science* 51(2): 364–381.

De Mesquita, Bruce Bueno, and Randolph M Siverson. 1995. "War and the survival of political leaders: A comparative study of regime types and political accountability." *American Political Science Review* 89(4): 841–855.

Dyer, Geoff. September 5, 2014. "Americans grow willing to accept military intervention." *Financial Times* .

Ellsberg, Daniel. 1968. The Theory and Practice of Blackmail. Technical report RAND. http://www.rand.org/pubs/papers/P3883/

Fearon, James. 1998. "Electoral Accountability and the Control of Politicians: Selecting Good types versus Sanctioning Poor Performance." In *Democracy, Accountability, and Representation*, eds. Susan C. Stokes Adam Przeworski, and Bernard Manin. Cambridge, UK: Cambridge University Press.

Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88(03): 577–592.

Fearon, James D. 1995. "Rationalist explanations for war." *International Organization* 49(3): 379–414.

Fischer, David Hackett. 1989. *Albion's Seed: Four British Folkways in America*. Oxford, UK: Oxford University Press.

Gadarian, Shana Kushner. 2010. "Foreign policy at the ballot box: how citizens use foreign policy to judge and choose candidates." *The Journal of Politics* 72(4): 1046–1062.

Gartner, Scott Sigmund. 2008. "The Multiple Effects of Casualties on Public Support for War: An Experimental Approach." *American Political Science Review* 102(1): 95–106.

Gartner, Scott Sigmund, and Gary M Segura. 1998. "War, Casualties, and Public Opinion." *Journal of Conflict Resolution* 42(3): 278–300.

Gelpi, Christopher, Peter D Feaver, and Jason Reifler. 2009*a*. *Paying the Human Costs of War*. Princeton, NJ: Princeton University Press.

Gelpi, Christopher, Peter D Feaver, and Jason Reifler. 2009*b*. *Paying the Human Costs of war: American Public opinion and Casualties in Military Conflicts*. Princeton, NJ: Princeton University Press.

George, Alexander. 1991. *Forceful Persuasion: Coercive Diplomacy as an Alternative to War.* Washington, DC: United States Institute of Peace.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W. W. Norton and Company.

Gershkoff, Amy, and Shana Kushner. 2005. "Shaping Public Opinion: The 9/11-Iraq Connection in the Bush Administration's Rhetoric." *Perspectives on Politics* 3(3): 525–537.

Gottfried, Matthew S, and Robert F Trager. 2016. "A preference for war: how fairness and rhetoric influence leadership incentives in crises." *International Studies Quarterly* 60(2): 243–257.

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. 2016. "On Provocation: Outrage, International Relations, and the Franco-Prussian War." *Security Studies* 26(1).

Hayashi, Yuka. February 1, 2015. "Hostages Drive Abe's Push to Boost Japan Military Action Overseas." *Financial Times* .

Hayes, Bradd C. 1989. *Naval Rules of Engagement: Management Tools for Crisis*. Santa Monica: Rand Corporation.

Hellman, Martin. 2013. "Avoiding Needless Wars, Part 5: Operation Northwoods." Unpublished paper, Defusing the Nuclear Threat Blog. https://nuclearrisk.wordpress.com/2013/03/30/avoiding-needless-wars-part-5-operation-northwoods/

Heneghan, Tom. November 13, 2015. "French president Hollande's televised address." *Reuters* .

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker. 2006. "Costly Punishment Across Human Societies." *Science* 312(5781): 1767–1770.

Herring, George C. 2008. *From Colony to Superpower: U.S. Foreign Relations Since 1776*. Oxford, UK: Oxford University Press.

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.

Jervis, Robert. 2013. "Back to Basics: State Power in a Contemporary World.". Oxford, UK: Oxford University Press.

Kagan, Donald. 1995. *On the Origins of War and the Preservation of Peace*. New York: Anchor Books, Doubleday.

Kalyvas, Stathis. 2006. *The Logic of Violence in Civil War*. Cambridge, UK: Cambridge University Press.

Klein, Christopher. 2012. "The Cuban Missile Crisis Pilot Whose Death May Have Saved Millions.".

Krebs, Ronald R., and Jennifer K. Lobasz. 2007. "Fixing the Meaning of 9/11 Hegemony, Coercion, and the Road to War in Iraq." *Security Studies* 16(3): 409–451.

Kurizaki, Shuhei. 2007. "Efficient secrecy: Public versus private threats in crisis diplomacy." *American Political Science Review* 101(03): 543–558.

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

Lai, Brian, and Dan Reiter. 2005. "Rally 'Round the Union Jack? Public Opinion and the Use of Force in the United Kingdom, 1948–2001." *International Studies Quarterly* 49(2): 255–272.

Larson, Eric V. 1996. *Casualties and Consensus: The Historical Role of Casualties in Domestic Support for U.S. Military Operations*. Santa Monica, CA: Rand Corporation.

Lebow, Richard Ned. 2008. *A Cultural Theory of International Relations*. New York: Cambridge University Press.

Liberman, Peter. 2013. "Retributive support for international punishment and torture." *Journal of Conflict Resolution* 57(2): 285–306.

Liberman, Peter. 2014. "War and torture as "just deserts"." *Public Opinion Quarterly* 78(1): 47–70.

Lichbach, Mark Irving. 1987. "Deterrence or escalation? The puzzle of aggregate studies of repression and dissent." *Journal of Conflict Resolution* 31(2): 266–297.

Lynn-Jones, Sean M. 1985. "A Quiet Success for Arms Control: Preventing Incidents at Sea." *International Security* 9(4): 154–184.

McCullough, Michael. 2008. *Beyond Revenge: The Evolution of the Forgiveness Instinct*. John Wiley & Sons.

Miller, William Ian. 2005. *Eye for an Eye*. Cambridge, UK: Cambridge University Press.

Miller, William Ian. 2009. *Bloodtaking and Peacemaking: Feud, Law, and Society in Saga Iceland*. Chicago, IL: University of Chicago Press.

Mueller, John E. 1973. *War, Presidents, and Public Opinion*. New York: John Wiley & Sons.

Myerson, Roger B. 2007. "Force and Restraint in Strategic Deterrence: a Game-Theorist's Perspective.".

Nisbett, Richard E., and Dov Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South*. Boulder, CO: Westview Press.

Nordstrom, Byron J. 2000. *Scandinavia since 1500*. Minneapolis, MN: University of Minnesota Press.

Nozick, Robert. 1969. "Coercion." In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, eds. Sidney Morgenbesser, Patrick Suppes, and Morton White. New York: St. Martin's Press.

O'Neill, Barry. 1999. *Honor, Symbols and War*. Ann Arbor, MI: University of Michigan Press.

Oren, Michael B. 2003. *Six Days of War: June 1967 and the Making of the Modern Middle East*. New York: Random House.

Oye, Kenneth A. 1993. *Economic discrimination and political exchange: World political economy in the 1930s and 1980s*. Princeton, NJ: Princeton University Press.

Pace, Julie. September 3, 2014. "Beheadings won't intimidate U.S., Obama says." *PBS Newshour/Associated Press* .

Pape, Robert A. 2003. "The Strategic Logic of Suicide Terrorism." *American Political Science Review* 97(3): 343–361.

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

Pickler, Nedra. January 7, 2015. "French Police Seal Off Villages in Hunt for Paris Killers." *Associated Press/Huffington Post* .

Pinker, Steven. 2011. *The Better Angels of Our Nature*. London: Penguin Group.

Powell, Robert. 1999. *In the Shadow of Power*. Princeton, NY: Princeton University Press.

Pruitt, Dean, Jeffrey Z Zubin, and Sung Hee Kim. 2004. *Social Conflict: Escalation, Stalemate and Settlement*. 3rd ed. McGraw-Hill Higher Education.

Rathbun, Brian, and Joshua D Kertzer. 2015. "Fair is Fair: Social Preferences and Reciprocity in International Politics." *World Politics* , 1–53.

Reed, John. February 4, 2015. "ISIS brutal killing of pilot hardens Jordan's resolve." *Financial Times* .

Rocco, Lorenzo, and Zié Ballo. 2008. "Provoking a civil war." *Public Choice* 134(3-4): 347–366.

Russett, Bruce. 1990. *Controlling the Sword*. Cambridge, UK: Cambridge University Press.

Sagan, Scott D., and Jeremi Suri. 2003. "The Madman Nuclear Alert: Secrey, Signaling, and Safety in October 1969." *International Security* 27(4): 150–183.

Sartori, Anne E. 2002. "The might of the pen: A reputational theory of communication in international disputes." *International Organization* 56(1): 121–149.

Schelling, Thomas C. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

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

Slantchev, Branislav L. 2011. *Military Threats: The Costs of Coercion and the Price of Peace*. Cambridge, UK: Cambridge Univeristy Press.

Staw, Barry M. 1981. "The escalation of commitment to a course of action." *Academy of management Review* 6(4): 577–587.

Stein, Rachel M. 2015. "War and Revenge: Explaining Conflict Initiation by Democracies." *American Political Science Review* 109(03): 556–573.

Stephens, Philip. 2014. "How to Avoid War in the East China Sea." *Financial Times* .

Stewart, Frank Henderson. 1994. *Honor*. Chicago, IL: University of Chicago Press.

Thatcher, Margaret. 1982. Hansard Debate. Technical report Hansard Debate, 14 April, Vol. 21, Col., 1,146.

TIME. December 15, 1941. "The U.S. At War: National Ordeal." *TIME* .

Tomz, Michael, Jessica Weeks, and Keren Yarhi-Milo. 2017. "How and Why Does Public Opinion Affect Foreign Policy in Democracies.".

Trager, Robert F. 2010. "Diplomatic Calculus in Anarchy: How Communication Matters." *The American Political Science Review* 104(02): 347–368.

Tucker, Robert W. 2007. *Woodrow Wilson and the Great War: Reconsidering America's Neutrality, 1914-1917*. Charlottesville, VA: University of Virginia Press.

Valentino, Benjamin A, Paul K Huth, and Sarah E Croco. 2010. "Bear Any Burden? How Democracies Minimize the Costs of War." *Journal of Politics* 72(02): 528.

Viscusi, Gregory, and Matthew Campbell. January 8, 2015. "French Police Seal Off Villages in Hunt for Paris Killers." *Bloomberg* .

Wawro, Geoffrey. 2005. *The Franco-Prussian War: The German Conquest of France in 1870-1871*. Cambridge, UK: Cambridge University Press.

Weisiger, Alex, and Keren Yarhi-Milo. 2015. "Revisiting Reputation: How Past Actions Matter in International Politics." *International Organization* 69(02): 473–495.

Weiss, Jessica, and Allan Dafoe. 2016. "Provocation, Public Opinion, and Crisis Escalation: Evidence from China.".

Weiss, Jessica Chen. 2013. "Nationalism and Escalation in the East China Sea." Unpublished paper, China Policy Institute Blog. https://cpianalysis.org/2013/11/14/nationalism-and-escalation-in-the-east-china-sea/

Coercion and Provocation Online Appendix

A Literature on Cost Sensitivity

As we discuss in our paper, the cost sensitivity assumption is essential to theories of public opinion and war (Larson 1996; Gelpi, Feaver, and Reifler 2009*a*). It is also essential to many other theories of war such as the balance of power theory (Slantchev 2011, 177), the bargaining model of war (Powell 1999, 111; Fearon 1995, 388) and some explanations for the democratic peace (Valentino, Huth, and Croco 2010, 529; Byman and Waxman 2002; Russett 1990, 46). Some specific examples of the cost sensitivity assumption in the literature on coercion and deterrence include:

"Coercion should work when the anticipated suffering associated with a threat exceeds the anticipated gains of defiance" (Byman and Waxman 2002, 10).

"The central task of coercive strategy [is] to create in the opponent the expectation of costs of sufficient magnitude to erode his motivation to continue what he is doing" (George 1991, 12).

"According to the standard rationalist narrative, coercion will ultimately succeed when the expected costs to the defender outstrip his interest in resisting the demand" (Abrahms 2013, 2).

"When the benefits that would be lost by concessions and the probability of attaining these benefits by continued resistance are exceeded by the costs of resistance and the probability of suffering these costs, the target concedes" (Pape 1996, 15–16).

"...if the initiator is deterrable (the threatened punishment exceeds his gains from attacking), he will see that an attack will make him worse off than restraint" (Achen and Snidal 1989, 151).

"...if the expected punishment exceeds the gain, then opponents will be deterred" (Achen and Snidal 1989, 152).

An advantage of the more general definition of coercion is that it fits better with the logic of reputation and honor. An agent's reputation for resolve and honor is harmed by both Schelling-coercion and brute-force. This distinction is thus not useful for our theory, and we prefer to avoid it by referring to both mechanisms as coercion.

B Survey Instrument

B.1 Text of the Vignette

B.1.1 Military Capability Treatment

- The United States and China disagree about the US's right to conduct military operations in the East China Sea.
- The US claims that international law grants it the right to conduct military operations in the international waters of the East China Sea. The US wants unrestricted access to these areas, stating that this is a matter of national interest.
- China claims that most of the East China Sea is within China's Air Defense Identification Zone (ADIZ) and that the US policy of conducting military operations in the East China Sea threatens China's sovereignty. China wants to restrict US access in the East China Sea.

- The dispute between the US and China has become tense, with both countries increasing their naval and air patrols in the East China Sea.
- **US Inferior.** Experts agree that the US's military capability is inferior to China's for a conflict in the East China Sea.
- **US Superior.** Experts agree that the US's military capability is superior to China's for a conflict in the East China Sea.

B.1.2 Provocation Treatment

Attack. Recently a Chinese military plane shot down an American military plane over the East China Sea. China claims the American plane was trespassing in China's ADIZ. The American pilot died and the plane was destroyed in the crash.

Collision. Recently there was a collision between an American and a Chinese military plane. The collision occurred because the Chinese plane was flying dangerous maneuvers around the American plane, making several close passes. On the third pass, the Chinese plane collided with the American plane. The American pilot died. The Chinese pilot just barely managed to eject and survive. Both planes were destroyed.

Non-Collision. Recently an American military plane and a Chinese military plane almost collided with each other. They almost collided because the Chinese plane was flying dangerous maneuvers around the American plane, making several close passes. No one was hurt.

Accident. Recently an American military plane that was conducting routine operations in the East China Sea crashed in a storm. The pilot died and the plane was destroyed.

Control. [No text.]

B.2 Text of the Survey

The text of the survey is presented below. We also display the variable names and coded values, which were not shown to the respondents.

Coercion Provocation Survey Experiment

Sample survey

Variable names are given in brackets. Coded values are given parentheses.

[InfCons] We are a research team at [redacted], studying foreign policy and international relations. In this study, you will be asked to read a short scenario describing an international crisis and then answer several questions. This survey should take approximately 5 minutes. Once you have completed the survey, you will receive $1.00. To be able to take the survey, you must live in the U.S. and use a non-mobile device. There are no known risks associated with this study. Although this study may not benefit you personally, we hope that our results will add to the knowledge about how international crises are perceived and operate. All of your responses will be held in confidence and we do not collect identifying information such as your name or email address. MTurk worker IDs will only be collected for the purposes of distributing compensation and will not be associated with survey responses. The results of this survey will be used for academic purposes only. If you have questions about your rights as a research participant, you may contact the [redacted] Human Subjects Committee: [redacted]. Additional information is available at: [redacted] Participation in this study is completely voluntary. You are free to decline to participate and to end participation at any time for any reason.

- ❍ I have read the above information and agree to participate in this study. (1)
- ❍ I do not agree to participate. (0)

Consider the following description of a hypothetical crisis the United States could face. We will ask for your opinion about how the US should handle a situation such as the one described

[USinferior] The United States and China disagree about the US's right to conduct military operations in the East China Sea.The US claims that international law grants it the right to conduct military operations in the international waters of the East China Sea. The US wants unrestricted access to these areas, stating that this is a matter of national interest. China claims that most of the East China Sea is within China's Air Defense Identification Zone (ADIZ) and that the US policy of conducting military operations in the East China Sea threatens China's sovereignty. China wants to restrict US access in the East China Sea. The dispute between the US and China has become tense, with both countries increasing their naval and air patrols in the East China Sea. Experts agree that the US's military capability is inferior to China's for a conflict in the East China Sea.

[Collision] Recently there was a collision between an American and a Chinese military plane. The collision occurred because the Chinese plane was flying dangerous maneuvers around the American plane, making several close passes. On the third pass, the Chinese plane collided with the American plane. The American pilot died. The Chinese pilot just barely managed to eject and survive. Both planes were destroyed.

Later in the survey, we will ask you some questions about the hypothetical scenario you just read. CLICK HERE to view the scenario in a new window so you can refer to it. We strongly recommend that you click on the link above.

Next, we are going to ask you some questions about the hypothetical scenario you just read.

[LevelEsc] In terms of using military force, how far should the US go to assert its claims in this dispute? Please select the maximum level of military force the US should use.

- ❍ The US should not assert its claims. (0)
- ❍ The US should assert its claims only if it can avoid using military force. (1)
- ❍ The US should assert its claims, using military force if necessary, but it should not risk war. (2)
- ❍ The US should assert its claims, using military force if necessary, even if this risks war. (3)

[EcCosts] In terms of future economic costs, how far should the US go to assert its claims in this dispute? Please select the maximum level of economic cost that you think the US should be willing to sustain. For some of the options, we provide real-world examples of costs in current dollars.

- The US should not assert its claims. (0
- The US should assert its claims only if it can avoid future economic costs. (1)
- $1 million (e.g., the monthly cost of flying an F-15 aircraft) (2)
- $10 million (e.g., the annual cost of flying an F15 aircraft) (3)
- $100 million (e.g., the cost of the US invasion of Grenada, 1983) (4)
- $1 billion (e.g., the cost of US military operations in Somalia, 1992-1994) (5)
- $10 billion (e.g., the cost of US military operations bombing ISIS over two years) (6)
- $100 billion (e.g., the cost to the US of the Gulf War, 1990-1991) (7)
- $1 trillion (e.g., the cost to the US of the Iraq War) (8)
- The US should assert its claims no matter the economic cost. (9)

[RiskWar] In terms of risking war with China, how far should the US go to assert its claims in this dispute? Please select the maximum level of risk of war with China the US should sustain.

- The US should not assert its claims. (0)
- The US should assert its claims only if it can avoid any risk of war with China. (1)
- The US should assert its claims only if it leads to at most a low risk of war with China. (2)
- The US should assert its claims only if it leads to at most a medium risk of war with China. (3)
- The US should assert its claims even if it leads to a high risk of war with China. (4)
- The US should assert its claims even if it leads to a very high risk of war with China. (5)
- The US should assert its claims no matter the risk of war with China. (6)

[MilDeaths] In terms of future US military fatalities, how far should the US go to assert its claims in this dispute? Please select the maximum number of US military fatalities the US should be willing to sustain. For some of the options, we provide real-world examples of military fatalities.

- The US should not assert its claims. (0)
- The US should assert its claims only if it can avoid US military fatalities. (1)
- 3 military fatalities (e.g., US military fatalities in 2014-present intervention against ISIS) (2)
- 30 military fatalities (e.g., US military fatalities in Somalia, 1992-1993) (3)
- 300 military fatalities (e.g., twice the number of US military fatalities in the Gulf War) (4)
- 3,000 military fatalities (e.g., US military fatalities in the Iraq War) (5)
- 30,000 military fatalities (e.g., US military fatalities in the Korean War or half the number of fatalities in the Vietnam War) (6)
- The US should assert its claims no matter the number of military fatalities. (7)

[Reputation] Do you think the events described in this scenario put the US's international reputation at stake?

- Yes (1)
- No (0)

[MagStake] Please tell us how much you think it is at stake.

- Not much at stake (0)
- Somewhat at stake (1)
- Very much at stake (2)

[MatCosts] How many deaths (US and Chinese, soldiers and civilians) do you think will occur as a result of this dispute in the coming year? Please select the number of deaths you think is most likely to occur.

- ❍ 0 deaths (0)
- ❍ About 1-24 deaths (1)
- ❍ About 25-200 deaths (2)
- ❍ More than 200 deaths (3)

[Coercion] Which of the following best describes China's behavior towards the US in the scenario?

- ❍ China wanted to avoid using force. (0)
- ❍ China was taking risks. (1)
- ❍ China was testing the US. (2)
- ❍ China was bullying the US. (3)

[MilCap] Given the information available in this scenario, which country do you think has superior military capabilities in the East China Sea: the US, China, or are they equal?

- ❍ The US probably has greater military capabilities than China. (1)
- ❍ China and the US probably have about equal military capabilities. (0)
- ❍ China probably has greater military capabilities than the US. (-1)

[MilInt] Given the information available in this scenario, how likely do think it is that China has plans to expand its military presence and capabilities in the East China Sea?

- ❍ Very unlikely (0)
- ❍ Unlikely (1)
- ❍ As likely as not (2)
- ❍ Likely (3)
- ❍ Very likely (4)

[RepCosts] In light of the events in this scenario, the US government abandoned its claim to having a right to conduct military operations in the East China Sea. How do you think the events described in this scenario will affect other countries' expectations of US behavior in future disputes?

- ❍ Other countries will think the US is less likely to stand firm in future disputes. (1)
- ❍ Other countries will not change their expectations about how likely the US is to stand firm in future disputes. (0)
- ❍ Other countries will think the US is more likely to stand firm in future disputes. (-1)

[MagRepBen] How much do you think the events described in this scenario will affect other countries' expectations of US behavior in future disputes?

- ❍ Other countries will think the US is moderately more likely to stand firm in future disputes. (0)
- ❍ Other countries will think the US is much more likely to stand firm in future disputes. (1)

[MagRepCost] How much do you think the events described in this scenario will affect other countries' expectations of US behavior in future disputes?

- ❍ Other countries will think the US is moderately less likely to stand firm in future disputes. (0)
- ❍ Other countries will think the US is much less likely to stand firm in future disputes. (1)

[ImpCons] Please tell us how important each of the following considerations was for you in deciding whether the US should maintain its claims in the East China Sea.

	Not important (0)	Slightly important (1)	Moderately Important (2)	Very Important (3)
[ImpCosts] Avoiding the risk of conflict with China that could cost the lives of Americans, Chinese and others.	❍	❍	❍	❍
[ImpStakes] Preserving US access to the East China Sea.	❍	❍	❍	❍
[ImpThreat] Resisting China's threat to US national interests and/or national security.	❍	❍	❍	❍
[ImpReputation] Maintaining the US's reputation.	❍	❍	❍	❍
[ImpHonor] Honoring the US pilot who died so that his death will not have been in vain.	❍	❍	❍	❍
[ImpRevenge] Punishing China.	❍	❍	❍	❍

Please tell us a bit about yourself.

[Gender] Are you male or female?
- ❍ Male (1)
- ❍ Female (2)
- ❍ I prefer not to answer (-98)

[Age] How old are you (in years)?

[Education] What is the highest level of education you have completed?
- ❍ Less than high school (1)
- ❍ High school or GED (2)
- ❍ Some college (3)
- ❍ 2-year college degree (4)
- ❍ 4-year college degree (5)
- ❍ Masters degree (6)
- ❍ Doctoral degree (7)
- ❍ Professional degree (e.g., JD or MD) (8)
- ❍ I prefer not to answer (-98)

[Race] What is your race?
- ❍ Caucasian (1)
- ❍ African-American (2)
- ❍ Asian (3)
- ❍ Hispanic (4)
- ❍ Native American (5)
- ❍ Pacific Islander (6)
- ❍ Other (7)
- ❍ I prefer not to answer (-98)

[USstate] In which US state do you currently reside?

- ❍ Alabama (1)
- ❍ Alaska (2)
- ❍ Arizona (3)
- ❍ Arkansas (4)
- ❍ California (5)
- ❍ Colorado (6)
- ❍ Connecticut (7)
- ❍ Delaware (8)
- ❍ District of Columbia (9)
- ❍ Florida (10)
- ❍ Georgia (11)
- ❍ Hawaii (12)
- ❍ Idaho (13)
- ❍ Illinois (14)
- ❍ Indiana (15)
- ❍ Iowa (16)
- ❍ Kansas (17)
- ❍ Kentucky (18)
- ❍ Louisiana (19)
- ❍ Maine (20)
- ❍ Maryland (21)
- ❍ Massachusetts (22)
- ❍ Michigan (23)
- ❍ Minnesota (24)
- ❍ Mississippi (25)
- ❍ Missouri (26)
- ❍ Montana (27)
- ❍ Nebraska (28)
- ❍ Nevada (29)
- ❍ New Hampshire (30)
- ❍ New Jersey (31)
- ❍ New Mexico (32)
- ❍ New York (33)
- ❍ North Carolina (34)
- ❍ North Dakota (35)
- ❍ Ohio (36)
- ❍ Oklahoma (37)
- ❍ Oregon (38)
- ❍ Pennsylvania (39)
- ❍ Puerto Rico (40)
- ❍ Rhode Island (41)
- ❍ South Carolina (42)
- ❍ South Dakota (43)
- ❍ Tennessee (44)
- ❍ Texas (45)
- ❍ Utah (46)
- ❍ Vermont (47)
- ❍ Virginia (48)
- ❍ Washington (49)
- ❍ West Virginia (50)
- ❍ Wisconsin (51)
- ❍ Wyoming (52)
- ❍ I do not reside in the United States (53)

[Income] What is your combined annual household income (in US dollars)?

- ❍ Less than $10,000 (1)
- ❍ $10,000-20,000 (2)
- ❍ $20,000-30,000 (3)
- ❍ $30,000-40,000 (4)
- ❍ $40,000-50,000 (5)
- ❍ $50,000-60,000 (6)
- ❍ $60,000-70,000 (7)
- ❍ $70,000-80,000 (8)
- ❍ $80,000-90,000 (9)
- ❍ $90,000-100,00 (10)
- ❍ More than $100,000 (11)
- ❍ I prefer not to answer (-98)

[Pol7scale] Below is a 7-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale, or haven't you thought much about this?

- ❍ 1. Extremely liberal (1)
- ❍ 2. Liberal (2)
- ❍ 3. Slightly liberal (3)
- ❍ 4. Moderate (4)
- ❍ 5. Slightly conservative (5)
- ❍ 6. Conservative (6)
- ❍ 7. Extremely conservative (7)
- ❍ Haven't thought much about this (8)
- ❍ I prefer not to answer (-98)

[PolParty] Generally speaking, do you think of yourself as a Democrat, a Republican or an Independent?
- ❍ Democrat (2)
- ❍ Republican (4)
- ❍ Independent (3)
- ❍ Haven't thought much about this (1)
- ❍ I prefer not to answer (-98)
- ❍ Other (-88) _____

[StrDem] Do you think of yourself as a strong Democrat?
- ❍ Yes (1)
- ❍ No (0)
- ❍ I prefer not to answer (-98)

[StrRep] Do you think of yourself as a strong Republican?
- ❍ Yes (1)
- ❍ No (0)
- ❍ I prefer not to answer (-98)

[DemRep] Do you think of yourself as closer to the Democratic Party or closer to the Republican Party?
- ❍ Closer to the Republican Party (2)
- ❍ Closer to the Democratic Party (1)
- ❍ Neither (4)
- ❍ I prefer not to answer (-98)

[MilExp] Have you ever served in the military?
- ❍ Yes (1)
- ❍ No (0)
- ❍ I prefer not to answer (-98)

[MilAssert] Please tell us whether you agree or disagree with the statements listed below.

	Strongly Disagree (1)	Disagree (2)	Neither Agree nor Disagree (3)	Agree (4)	Strongly Agree (5)
[MilAssert_1] The best way to ensure world peace is through American military strength.	❍	❍	❍	❍	❍
[MilAssert_2] The use of military force only makes problems worse.	❍	❍	❍	❍	❍
[MilAssert_3] Going to war is unfortunate, but sometimes the only solution to international problems.	❍	❍	❍	❍	❍

[ReadForei] How frequently do you read about foreign affairs, international relations, or international history in the news, in articles, or in books?

- Everyday (1)
- 2-4 times a week (2)
- 2-4 times a month (3)
- 5-20 times a year (4)
- Less than 5 times a year (5)
- I prefer not to answer (-98)

[OnlineDisc] Do you participate in any online discussions of MTurk HITs? Your answer to this question will not affect the approval of this HIT.

- Yes (1)
- No (0)
- I prefer not to answer (-98)

[SimSurvey] In addition to this study, approximately how many other studies related to foreign policy or international relations have you completed on MTurk in the past year? Your answer to this question will not affect the approval of this HIT.

- None (0)
- 1-5 (1)
- 6-10 (4)
- 11-15 (5)
- More than 15 (6)
- I prefer not to answer (-98)

[CatchTri] Have you ever died while watching television?

- Always (5)
- Most of the time (4)
- Sometimes (3)
- Rarely (2)
- Never (1)

Thank you for your participation!

C Additional Analyses

C.1 Balance Checks

We check if the background characteristics of the respondents are balanced across treatment assignments. In Table 3, we provide summary statistics of respondents' background characteristics. Table 4 shows that the background characteristics of respondents do not individually or jointly predict treatment assignment.

Table 3: Summary Statistics of Respondents' Background Characteristics

Variables	Attack	Collision	Non-Collision	Accident	Control
Mean age	34.97	35.74	35.48	35.77	35.28
Prop. male	0.53	0.54	0.55	0.55	0.50
Prop. with college degree	0.53	0.52	0.51	0.50	0.52
Prop. white	0.81	0.79	0.76	0.79	0.82
Prop. income $>$ \$50K	0.50	0.48	0.53	0.50	0.48
Mean PID 7	3.47	3.37	3.55	3.48	3.46
Prop. Democrat	0.42	0.43	0.41	0.40	0.39
N	350	380	352	322	357

PID 7 is a 7-point political ideology scale measure, in which 1 is extremely liberal and 7 is extremely conservative.

Table 4: Balance on Background Characteristics

	Accident vs. Control	Collision vs. Non-Collision	Attack vs. Control
Age	0.017	0.002	−0.006
	(0.020)	(0.018)	(0.020)
Male	0.028	−0.007	0.014
	(0.019)	(0.019)	(0.019)
College Degree	−0.012	0.008	0.006
	(0.020)	(0.019)	(0.019)
White	−0.017	0.022	0.0002
	(0.020)	(0.018)	(0.020)
Income > $50K	0.009	−0.029	0.008
	(0.020)	(0.019)	(0.019)
PID 7	−0.002	−0.025	0.016
	(0.023)	(0.023)	(0.022)
Democrat	0.005	−0.005	0.027
	(0.023)	(0.022)	(0.023)
Constant	0.475***	0.519***	0.494***
	(0.019)	(0.019)	(0.019)
N	678	731	705
F Statistic	0.554 (df = 7; 670)	0.932 (df = 7; 723)	0.304 (df = 7; 697)

For each vignette design comparison, we regress the treatment assignment on the seven background characteristics. The regressors have been mean-standardized for ease of interpretation. Heteroskedasticity-robust standard errors are reported below coefficients. All *p*-values reported are two-sided.

C.2 Reputation and Information Mediators

Our study established that fatalities experienced in a coercive context can provoke (increase resolve), even when the fatality was inadvertent; it does not provoke when the fatality was an accident outside of the immediate coercive context; information revelation is not able to fully account for this result.

We explore potential mechanisms for this effect by *ruling out potential mediators*[1]. We gain evidence against a potential mediator if the effect of the treatment on the mediator for every observation appears to be zero. On the other hand, if these potential mediators are affected by treatment in the theorized direction then we gain evidence that they do in fact mediate the treatment effect.[2] This approach to mediation analysis stays closer to the experimental design than typical approaches,

[1] Gerber and Green 2012, §10.4

[2] For each comparison and potential mediator, we conduct two tests using randomization inference as recommended by Gerber and Green 2012, §10.4. First, we test the sharp null of no effect

which we recommend for more credible causal inference whenever it is possible. In particular, it does not require that we are able to model away the endogeneity between the mediator and the outcome.[3]

We first conduct a confirmatory test of whether it is more likely that "reputation appears to be at stake" in our provocation conditions. We use several synonyms for this concept, including prestige, international reputation, image abroad, credibility, honor, national honor, and international status. Combining all of these, we find some evidence that reputation is more at stake in the Attack vs. Control conditions (for dichotomous and cardinal one-tailed comparisons, $p = 0.12$ and $p = 0.06$) but not in Collision vs. No-Collision conditions (Figure 6). Looking at our specific terms, recognizing that this is exploratory, we see suggestive evidence that Attack vs Control puts "Prestige," "Credibility," and especially "National Honor" at stake; Collision vs. No-Collision seems to put "International Reputation" at stake (see Figure 7).

As additional exploratory mediation analysis, we asked "how important each of" a set of specific considerations was for the respondent. The considerations included the costs of conflict ("Avoiding the risk of conflict with China that could cost the lives of Americans, Chinese and others"), the stakes ("preserving US access to the East China Sea"), the broader threat from China ("resisting China's threat to US national interests and/or national security"), reputation ("maintaining the US's reputation"), an honor debt ("honoring the US pilot so that his death will not have been in vain"), and vengeance ("punishing China"). This analysis reveals, shown in Figure 5, several things.[4] While "punishing China" was overall not considered to be an important consideration for most respondents, Attack vs. Control, and Collision vs. No-Collision both increased the importance by a large magnitude (two-sided $p < 0.001$), and Attack vs. Control did so by almost twice as much. We also see clear evidence that considerations of an honor debt ("honoring the US pilot") are greater for Attack and Collision, compared with Accident (we did not ask about this under Control or Non-Collision, since there was no death). Attack vs. Control, and Collision vs. No-Collision, both reduced the importance of the costs of conflict (two-sided $p < 0.01$), which was overall the most important consideration. To a weaker extent, we saw some evidence of reputation responding to Collision vs. Non-Collision, and broader threat responding to Attack vs. Control.

Overall this analysis provides suggestive evidence of the importance of vengeance, honor, and reputational considerations in motivating provocation, facilitated by the attenuation of the dominat-

---

for all subjects using the difference-in-means. Nevertheless, an average effect of zero according to this first test could be caused by the positive effects experienced by some subjects canceling out the negative effects experienced by other subjects. To gain evidence about whether heterogeneous effects could be driving this result, we then test the null hypothesis of homogeneous effects using the difference-in-variances between the treatment condition and control condition as the test statistic.

[3] Bullock, Green, and Ha 2010.

[4] Table 7 contains the results of the randomization inference analysis using the difference-of-means as the test statistic. Table 8 contains the results of the randomization inference analysis using the difference-of-variances as the test statistic.

ing consideration of avoiding costly conflict. The increase in respondents' support for escalating the conflict seems to be driven more by concern for revenge and honor than by concern for the U.S.'s deterrence reputation. The treatments that involved a U.S. pilot fatality strongly induced respondents to support punishing China and honoring the U.S. pilot who died. Our results are consistent with past empirical works.[5] that suggest that preference for retribution against the enemy can increase states' and voters' support for using military force.

[5]Stein 2015; Liberman 2013, 2014.

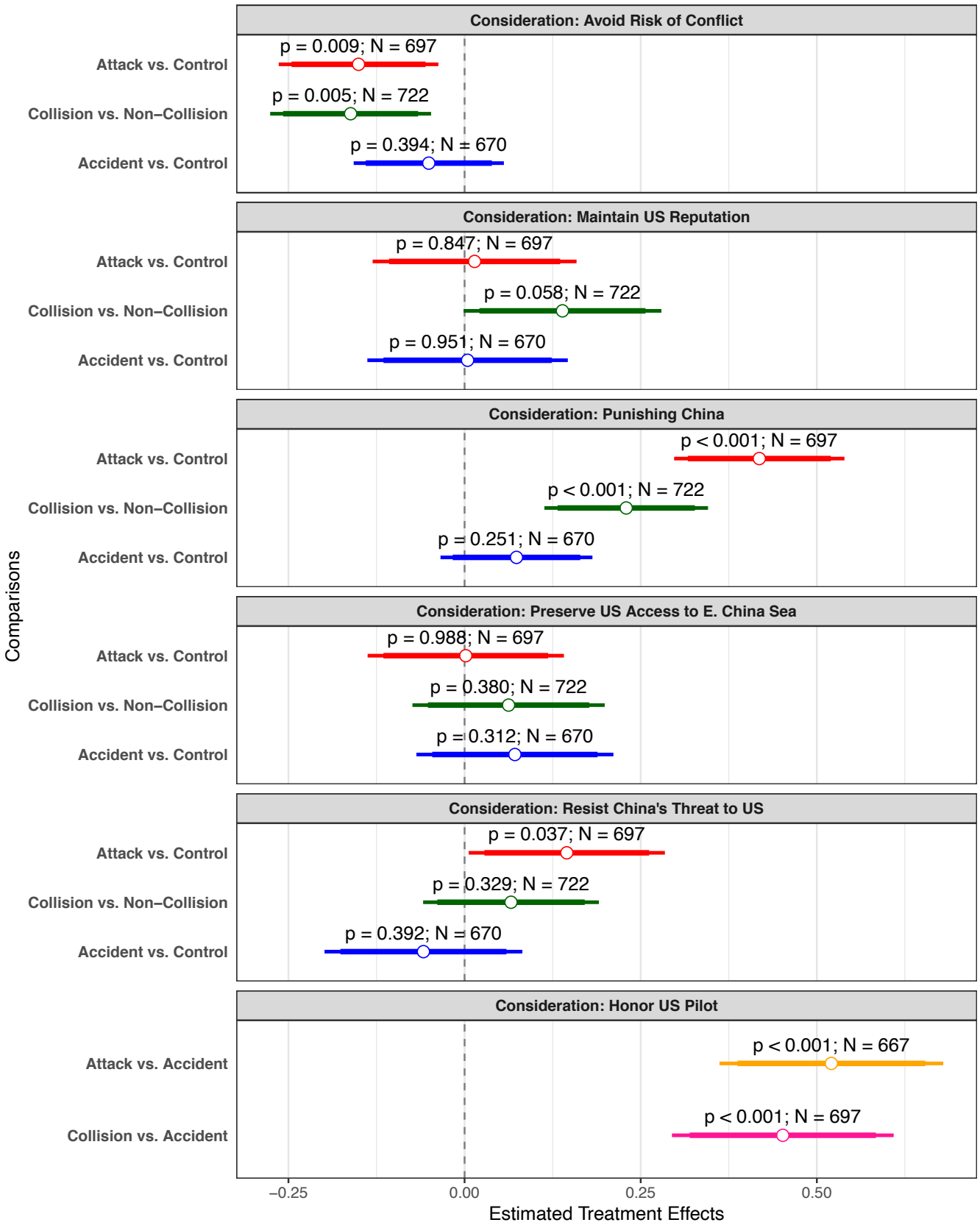
Consideration: Avoid Risk of Conflict
Attack vs. Control
p = 0.009; N = 697
Collision vs. Non−Collision
p = 0.005; N = 722
Accident vs. Control
p = 0.394; N = 670
Consideration: Maintain US Reputation
Attack vs. Control
p = 0.847; N = 697
Collision vs. Non−Collision
p = 0.058; N = 722
Accident vs. Control
p = 0.951; N = 670
Consideration: Punishing China
Attack vs. Control
p < 0.001; N = 697
Collision vs. Non−Collision
p < 0.001; N = 722
Accident vs. Control
p = 0.251; N = 670
Consideration: Preserve US Access to E. China Sea
Attack vs. Control
p = 0.988; N = 697
Collision vs. Non−Collision
p = 0.380; N = 722
Accident vs. Control
p = 0.312; N = 670
Consideration: Resist China's Threat to US
Attack vs. Control
p = 0.037; N = 697
Collision vs. Non−Collision
p = 0.329; N = 722
Accident vs. Control
p = 0.392; N = 670
Consideration: Honor US Pilot
Attack vs. Accident
p < 0.001; N = 667
Collision vs. Accident
p < 0.001; N = 697
Comparisons
−0.25
0.00
0.25
0.50
Estimated Treatment Effects

Figure 5: Respondents' Considerations When Deciding to Maintain Claims. We estimate the effects of the scenarios on respondents' considerations when deciding to support the US's maintaining it claims. The two-sided *p*-values are calculated from randomization inference using the difference-in-means as the test-statistic; we generate 10000 simulations of placebo treatment statuses in our randomization inference. The mean response for each experimental group and consideration is reported in Figure 8. Note for "Honor US Pilot," we compare the Attack condition with the Accident condition and the Collision condition with the Accident condition. We do so because respondents assigned to the Control and Non-Collision conditions were not given "Honor US Pilot" as a consideration since no US pilot died in those conditions' vignettes.

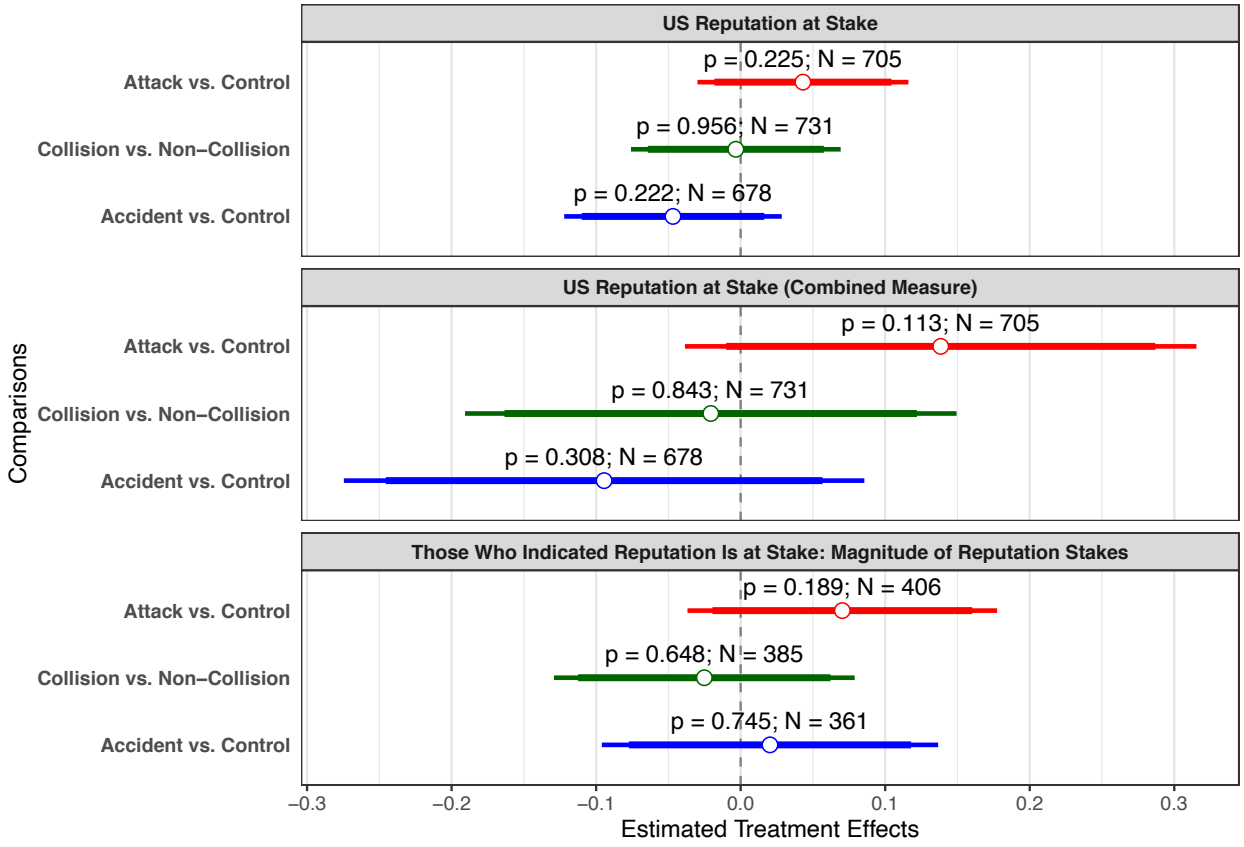

Figure 6: Reputation at Stake. The two-sided *p*-values are calculated from randomization inference using the difference-in-means as the test-statistic; we generate 10000 simulations of placebo treatment statuses in our randomization inference. The variable **US Reputation at Stake** is coded 1 if the respondent indicated that the type of the reputation she was assigned was at stake and 0 otherwise (see Figure 7). If the respondent indicated the type of reputation is at stake, she is asked how much reputation is at stake. For the variable **US Reputation at Stake (Combined Measure)**, 0 means not at stake, 1 means not much at stake, 2 means somewhat at stake, and 3 means very much at stake. For those who indicated the reputation is at stake, we compare the magnitude of the reputation at stake by treatment assignment.

Table 5: Reputation at Stake (Difference-of-Means). We evaluate reputation concerns as potential mediators. To rule out a mediator, we test whether the treatment had zero effect on the mediator for every subject. To test the sharp null of no effect for every subject, we perform randomization inference using difference-of-means as the test statistic. This table reports the estimated differences-in-means along with the two-sided p-values; we generate 10000 simulations of placebo treatment statuses in our randomization inference.

Considerations	Comparisons	Estimates	p-values
US Reputation at Stake	Accident vs. Control	-0.045	0.222
US Reputation at Stake	Attack vs. Control	0.045	0.225
US Reputation at Stake	Collision vs. Non-Collision	-0.002	0.956
US Reputation at Stake (Combined Measure)	Accident vs. Control	-0.092	0.308
US Reputation at Stake (Combined Measure)	Attack vs. Control	0.142	0.113
US Reputation at Stake (Combined Measure)	Collision vs. Non-Collision	-0.018	0.843
Those Who Indicated Reputation Is at Stake: Magnitude of Reputation Stakes	Accident vs. Control	0.010	0.745
Those Who Indicated Reputation Is at Stake: Magnitude of Reputation Stakes	Attack vs. Control	0.040	0.189
Those Who Indicated Reputation Is at Stake: Magnitude of Reputation Stakes	Collision vs. Non-Collision	-0.014	0.648

Table 6: Reputation at Stake (Difference-of-Variances). We evaluate reputation concerns as potential mediators. To rule out a mediator, we test whether the treatment had zero effect on the mediator for every subject. To test the sharp null of homogeneous treatment effects, we perform randomization inference using difference-of-variances as the test statistic. This table reports the estimated differences-in-variances along with the two-sided p-values; we generate 10000 simulations of placebo treatment statuses in our randomization inference.

Considerations	Comparisons	Estimates	p-values
US Reputation at Stake	Accident vs. Control	0.003	0.352
US Reputation at Stake	Attack vs. Control	-0.007	0.054
US Reputation at Stake	Collision vs. Non-Collision	0.0001	0.983
US Reputation at Stake (Combined Measure)	Accident vs. Control	0.022	0.739
US Reputation at Stake (Combined Measure)	Attack vs. Control	0.038	0.575
US Reputation at Stake (Combined Measure)	Collision vs. Non-Collision	-0.035	0.599
Those Who Indicated Reputation Is at Stake: Magnitude of Reputation Stakes	Accident vs. Control	-0.014	0.542
Those Who Indicated Reputation Is at Stake: Magnitude of Reputation Stakes	Attack vs. Control	0.0004	0.986
Those Who Indicated Reputation Is at Stake: Magnitude of Reputation Stakes	Collision vs. Non-Collision	-0.006	0.798

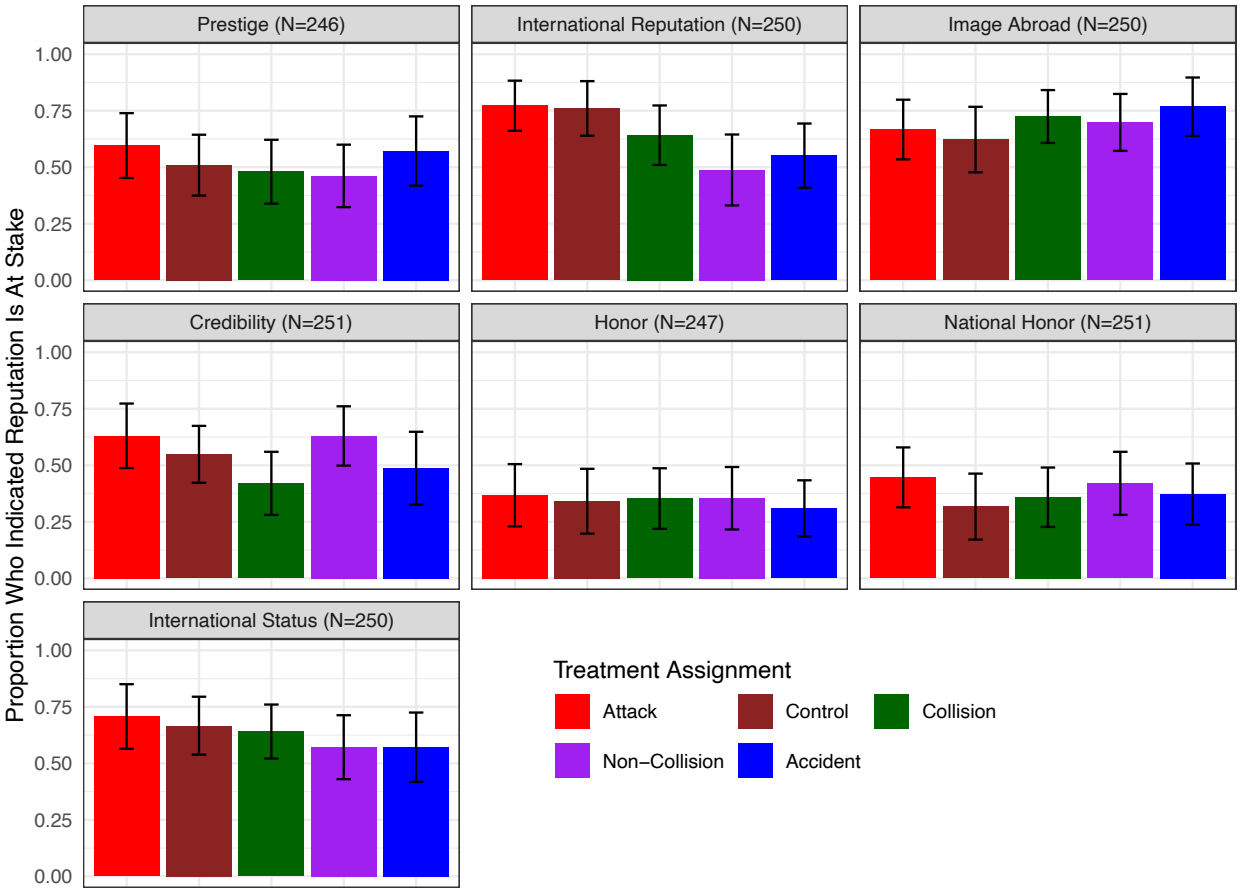

Figure 7: Types of Reputation at Stake. Each respondent is randomly assigned one of the seven reputation types. Respondents are asked whether they think the reputation type assigned to them is at stake in the dispute. The error bars include the 95 percent confidence interval for each estimate.

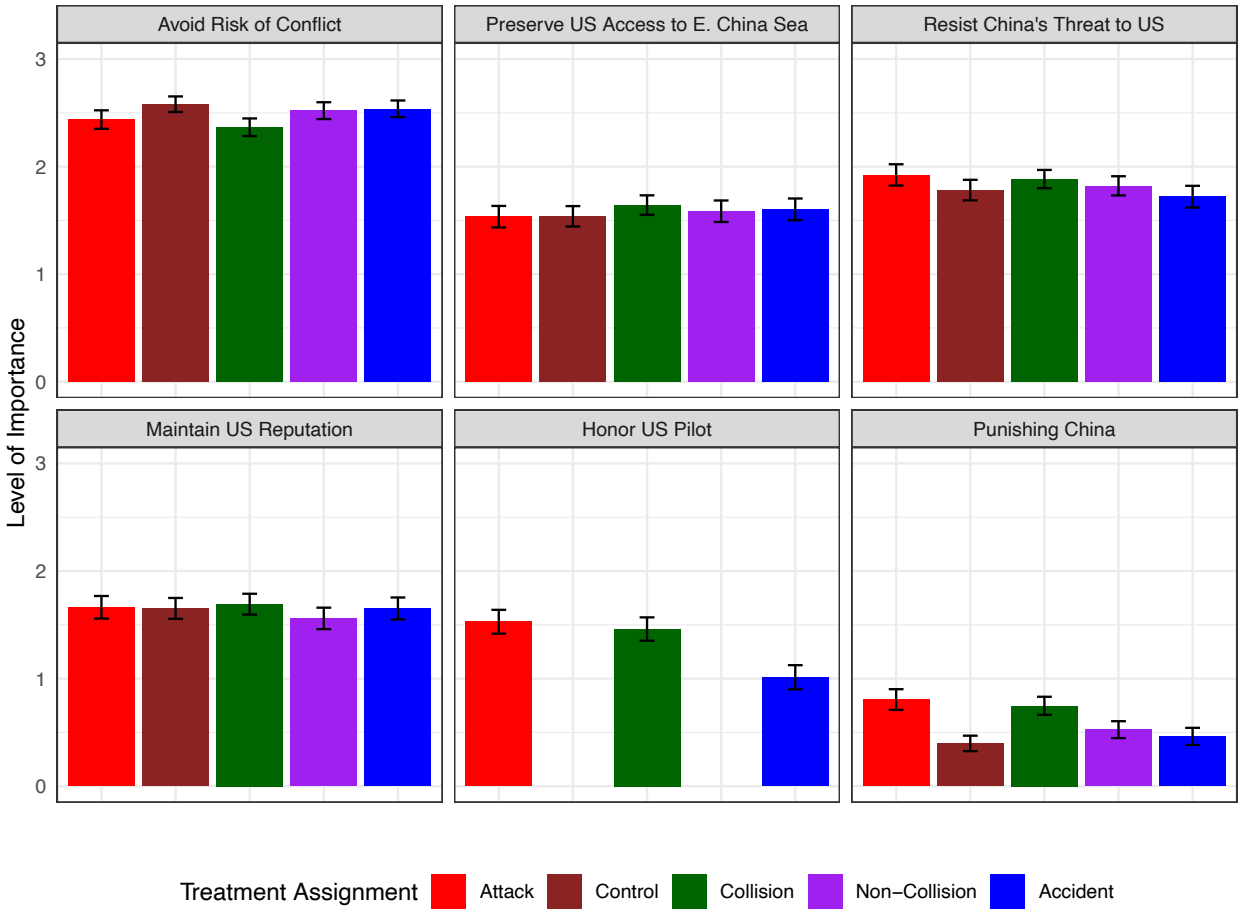

Figure 8: Respondents' Considerations When Deciding to Maintain Claims. Respondents were given a list of considerations and asked how important each one is for their decision to support the US's maintaining its claims. The consideration **Honoring US Pilot** is only shown to respondents assigned to read vignettes in which the American pilot died. For each consideration, 0 means not important at all, 1 means slightly important, 2 means moderately important, and 3 means very important.

Table 7: Respondents' Considerations When Deciding to Maintain Claims (Difference-of-Means). Each consideration is a potential mediator. To rule out a mediator, we test whether the treatment had zero effect on the mediator for every subject. To test the sharp null of no effect for every subject, we perform randomization inference using difference-of-means as the test statistic. This table reports the estimated differences-in-means along with the two-sided *p*-values; we generate 10000 simulations of placebo treatment statuses in our randomization inference.

Considerations	Comparisons	Estimates	*p*-values
Avoid Risk of Conflict	Accident vs. Control	-0.049	0.394
Avoid Risk of Conflict	Attack vs. Control	-0.148	0.009
Avoid Risk of Conflict	Collision vs. Non-Collision	-0.160	0.005
Maintain US Reputation	Accident vs. Control	0.004	0.951
Maintain US Reputation	Attack vs. Control	0.014	0.847
Maintain US Reputation	Collision vs. Non-Collision	0.137	0.058
Punishing China	Accident vs. Control	0.071	0.251
Punishing China	Attack vs. Control	0.413	<0.001
Punishing China	Collision vs. Non-Collision	0.227	<0.001
Preserve US Access to E. China Sea	Accident vs. Control	0.070	0.312
Preserve US Access to E. China Sea	Attack vs. Control	0.001	0.988
Preserve US Access to E. China Sea	Collision vs. Non-Collision	0.062	0.380
Resist China's Threat to US	Accident vs. Control	-0.059	0.392
Resist China's Threat to US	Attack vs. Control	0.143	0.037
Resist China's Threat to US	Collision vs. Non-Collision	0.065	0.329
Honor US Pilot	Attack vs. Accident	0.517	<0.001
Honor US Pilot	Collision vs. Accident	0.449	<0.001

Table 8: Respondents' Considerations When Deciding to Maintain Claims (Difference-of-Variances). Each consideration is a potential mediator. To rule out a mediator, we test whether the treatment had zero effect on the mediator for every subject. To test the sharp null of homogeneous treatment effects, we perform randomization inference using difference-of-variances as the test statistic. This table reports the estimated differences-in-variances along with the two-sided p-values; we generate 10000 simulations of placebo treatment statuses in our randomization inference.

Considerations	Comparisons	Estimates	p-values
Avoid Risk of Conflict	Accident vs. Control	0.017	0.827
Avoid Risk of Conflict	Attack vs. Control	0.190	0.018
Avoid Risk of Conflict	Collision vs. Non-Collision	0.107	0.182
Maintain US Reputation	Accident vs. Control	-0.002	0.971
Maintain US Reputation	Attack vs. Control	0.132	0.070
Maintain US Reputation	Collision vs. Non-Collision	0.016	0.818
Punishing China	Accident vs. Control	0.058	0.468
Punishing China	Attack vs. Control	0.369	<0.001
Punishing China	Collision vs. Non-Collision	0.137	0.091
Preserve US Access to E. China Sea	Accident vs. Control	0.021	0.773
Preserve US Access to E. China Sea	Attack vs. Control	0.077	0.278
Preserve US Access to E. China Sea	Collision vs. Non-Collision	-0.097	0.161
Resist China's Threat to US	Accident vs. Control	0.005	0.943
Resist China's Threat to US	Attack vs. Control	0.046	0.509
Resist China's Threat to US	Collision vs. Non-Collision	-0.018	0.790
Honor US Pilot	Attack vs. Accident	0.053	0.472
Honor US Pilot	Collision vs. Accident	0.106	0.151

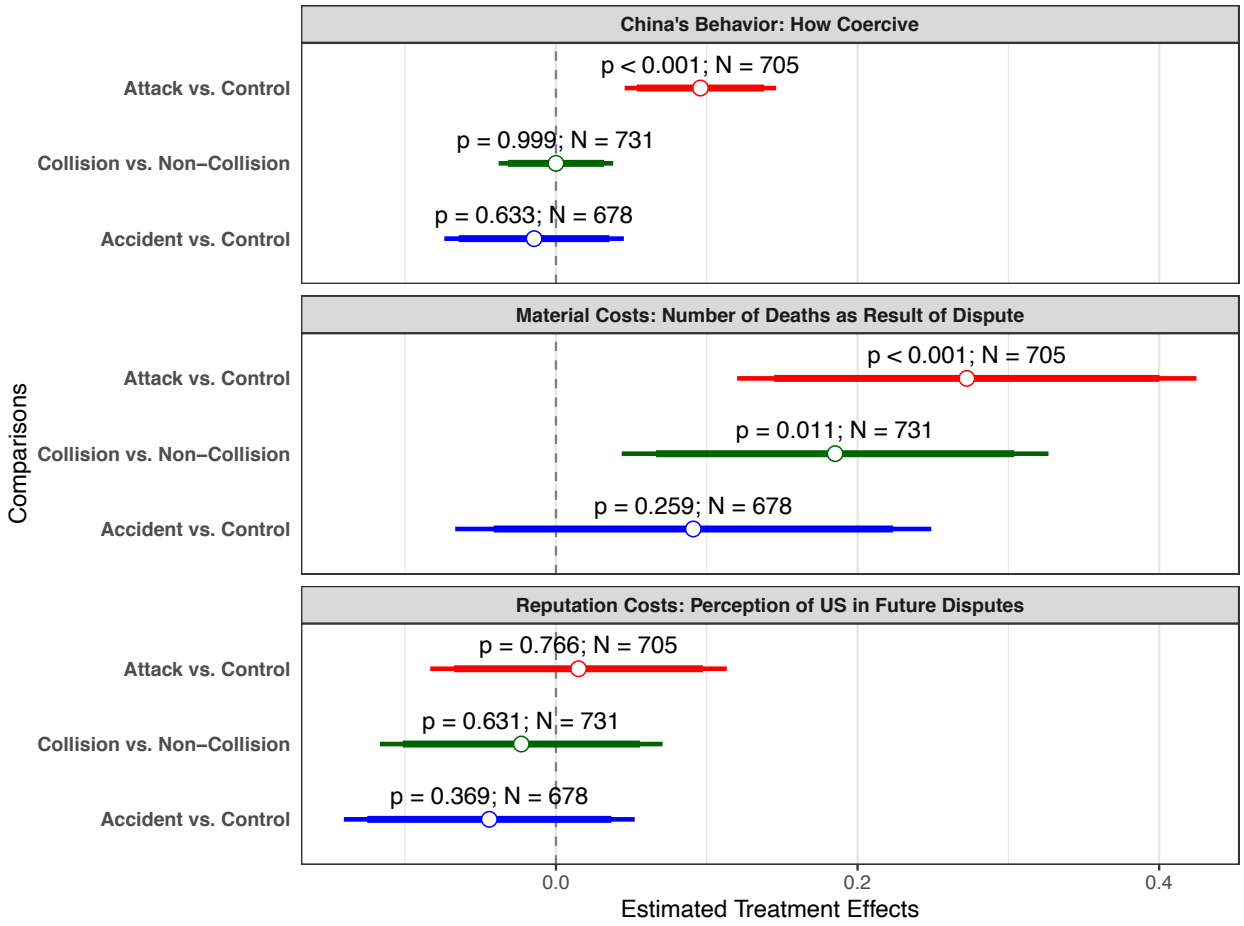

Figure 9: Additional Outcomes. We estimate the effects of the scenarios on additional outcomes related to coercion including: 1) how coercive China is perceived to be in the scenario, 2) the number of deaths that respondents predict will result from the dispute, and 3) perceptions of the US's reputation for resolve in the future. Perceptions of China's coercive behavior are measured on 4-point scale such that higher values represent higher levels of coercion. Respondents' estimates of deaths that could result from the dispute in the upcoming year are measured on a 4-point scale such that 0 means 0 deaths, 1 means 1-24 deaths, 2 means 25-200 deaths, and 3 means more than 200 deaths. Respondents' perceptions of US reputation are measured on a 3-point scale; -1 means that they perceive US reputation for resolve will suffer, 0 means that they perceive US reputation for resolve will remain unchanged, and 1 means that they perceive US reputation for resolve will improve.

C.3 Additional Figures

The following sections present additional figures not included in the article. For each coefficient plot, we display the 95 and 99 percent heteroskedasticity-robust confidence intervals unless specified.

C.3.1 Main Outcomes

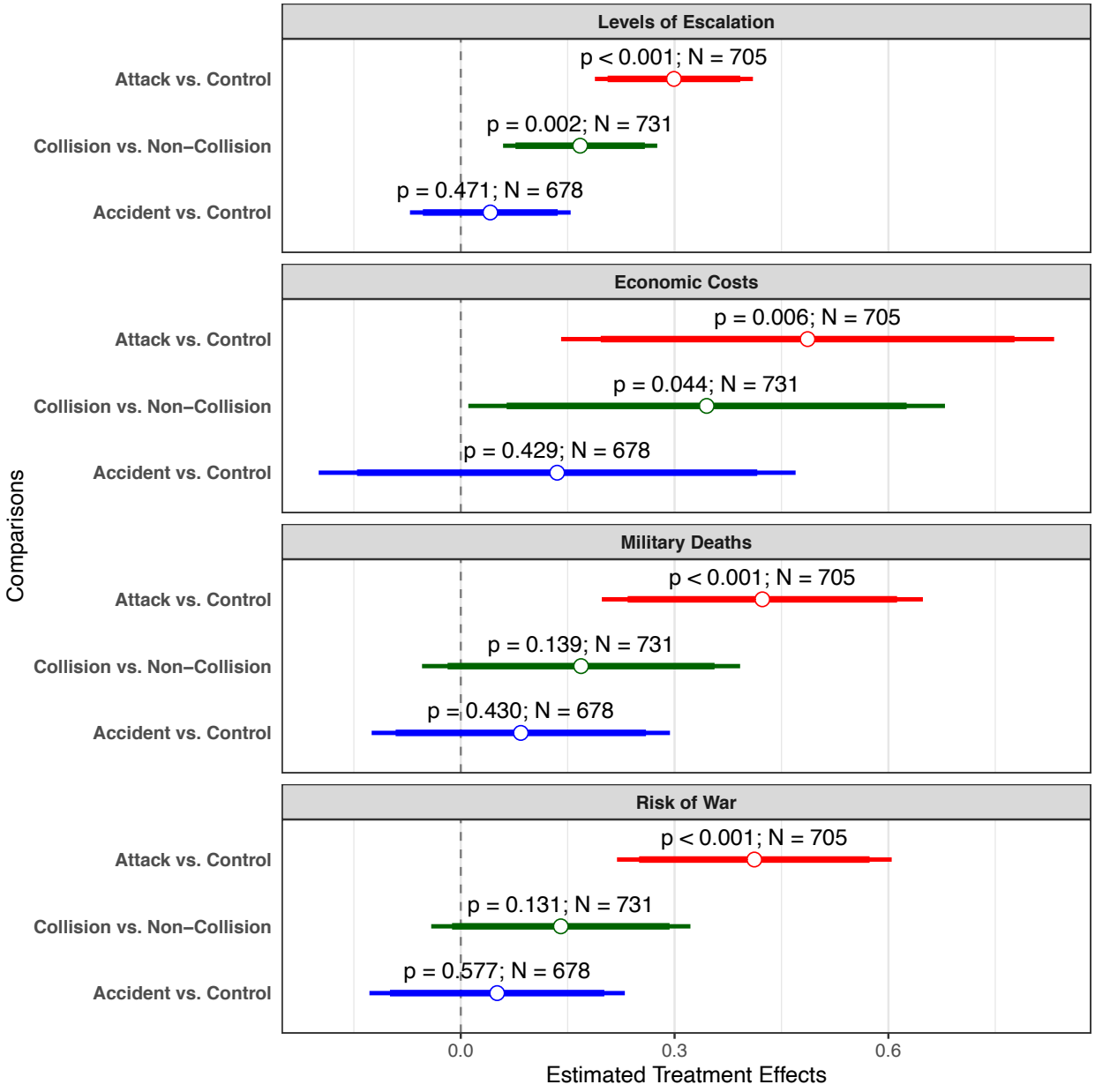

Figure 10: Main Outcomes: Escalation in the Dispute. The variable **Levels of Escalation** ranges from 0 to 3. The variable **Economic Costs** ranges from 0 to 10. The variable **Military Deaths** ranges from from 0 to 7. The variable **Risk of War** ranges from 0 to 6. For all variables, larger values means greater levels of escalation.

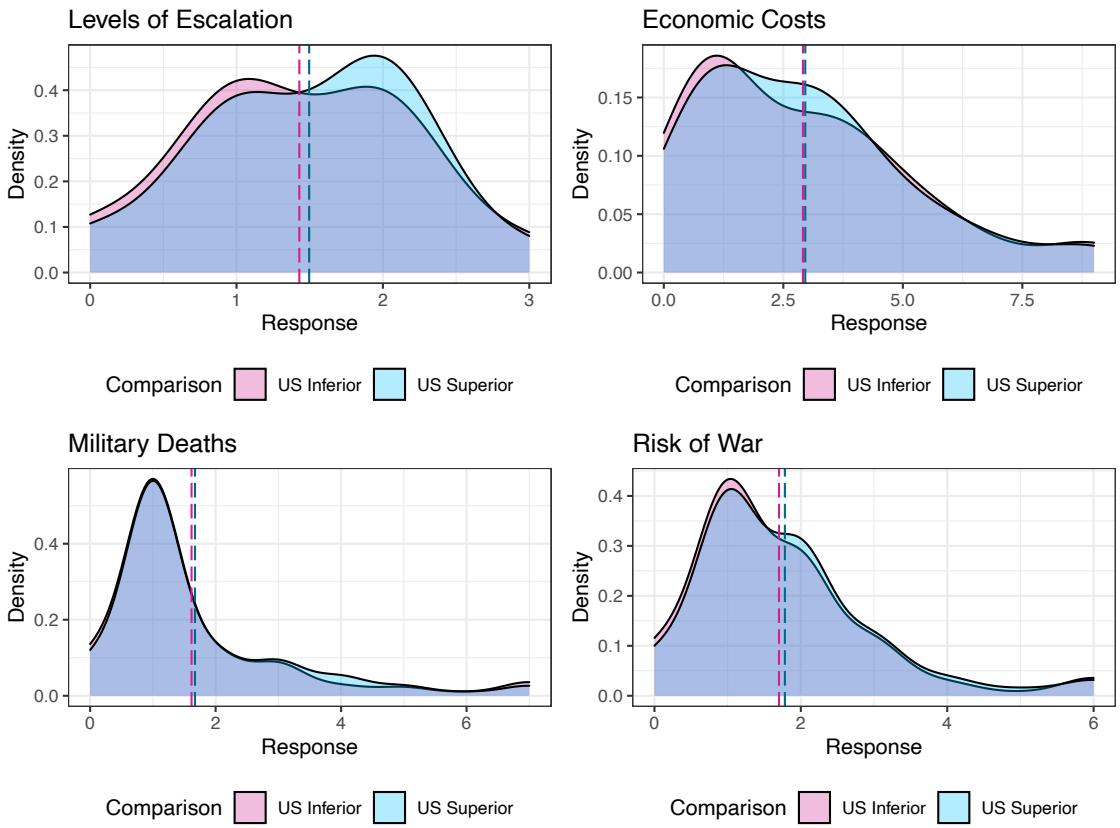

Figure 11: Military Capability Treatment: Distribution of Responses. We experimentally manipulated the introductory text of the scenario. Each respondent has a 50 percent chance of being told that "experts agree that the US's military capability is inferior to China's for a conflict in the East China Sea" and a 50 percent chance of being told that "experts agree that the US's military capability is superior to China's for a conflict in the East China Sea." We estimate the effects of this experimental manipulation on the four main outcomes. The graph shows the distributions of responses by treatment group and outcome measure. The dashed lines are the mean responses for each treatment assignment.

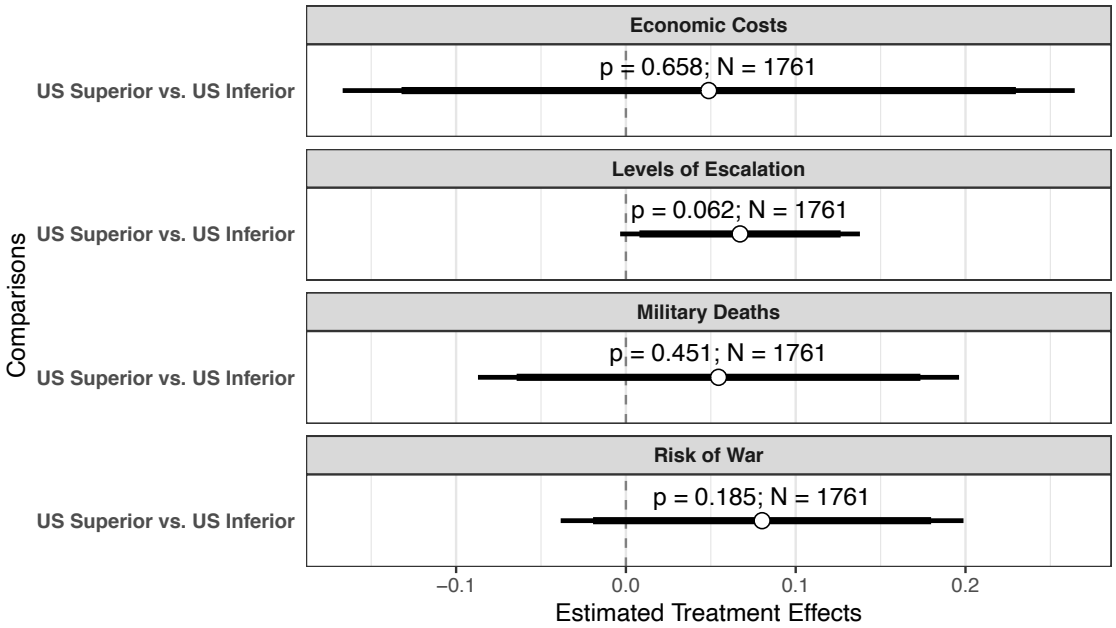

Figure 12: Effects of Military Capability Treatment. We experimentally manipulated the introductory text of the scenario. Each respondent has a 50 percent chance of being told that "experts agree that the US's military capability is inferior to China's for a conflict in the East China Sea" and a 50 percent chance of being told that "experts agree that the US's military capability is superior to China's for a conflict in the East China Sea." We estimate the effects of this experimental manipulation on the four main outcomes. The graph shows the estimated difference-in-means (US Superior - US Inferior) for each outcome measure. The results above suggest that the military capability manipulation did not affect the four main outcomes.

C.4 Main Outcomes: Distribution of Responses

In the next three sets of figures, we plot the distribution of responses to the four main outcome measures. For each treatment group and outcome, we also plot the mean response along with its 95 percent confidence interval.

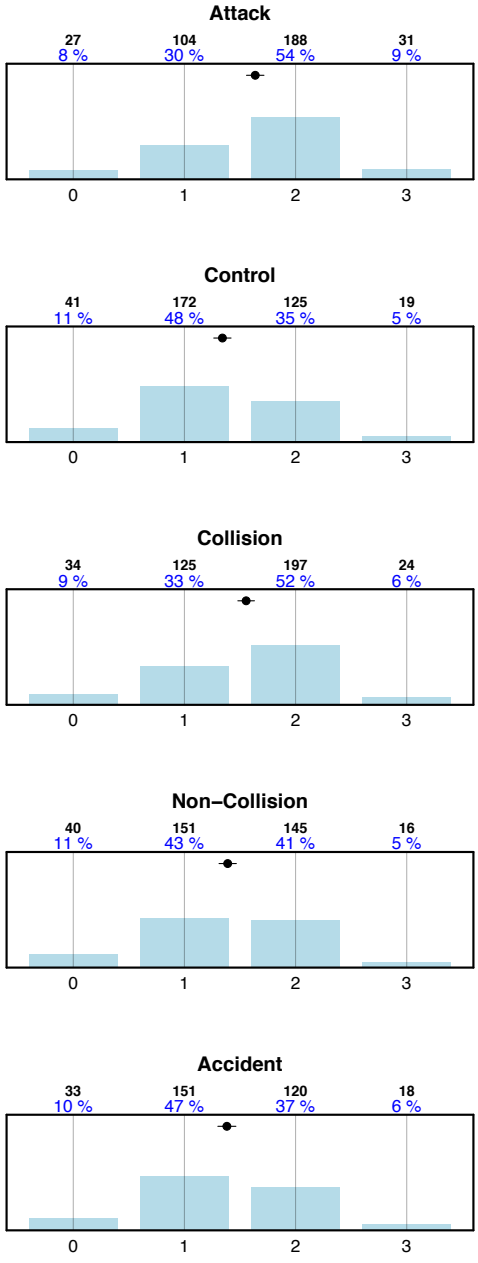

Figure 13: Distribution of Responses: Levels of Escalation. For the **Levels of Escalation** variable, 0 means the US should not assert its claims; 1 means the US should assert its claims only if it can avoid using military force; 2 means the US should assert its claims, using military force if necessary, but it should not risk war; 3 means the US should assert its claims, using military force if necessary, even if this risks war.

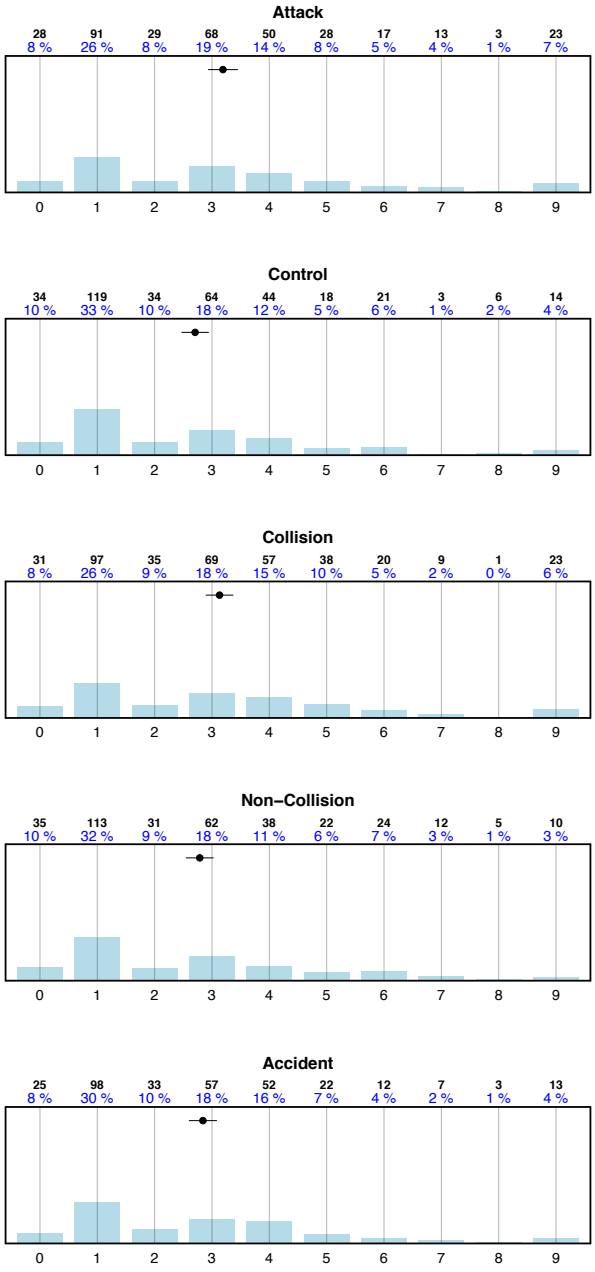

Figure 14: Distribution of Responses: Economic Costs. For the **Economic Costs** variable, 0 means the US should not assert its claims; 1 means the US should assert its claims only if it can avoid future economic costs; 2 means $1 million; 3 means $10 million; 4 means $100 million; 5 means $1 billion; 6 means $10 billion; 7 means $100 billion; 8 means $1 trillion; 9 means the US should assert its claims no matter the economic cost.

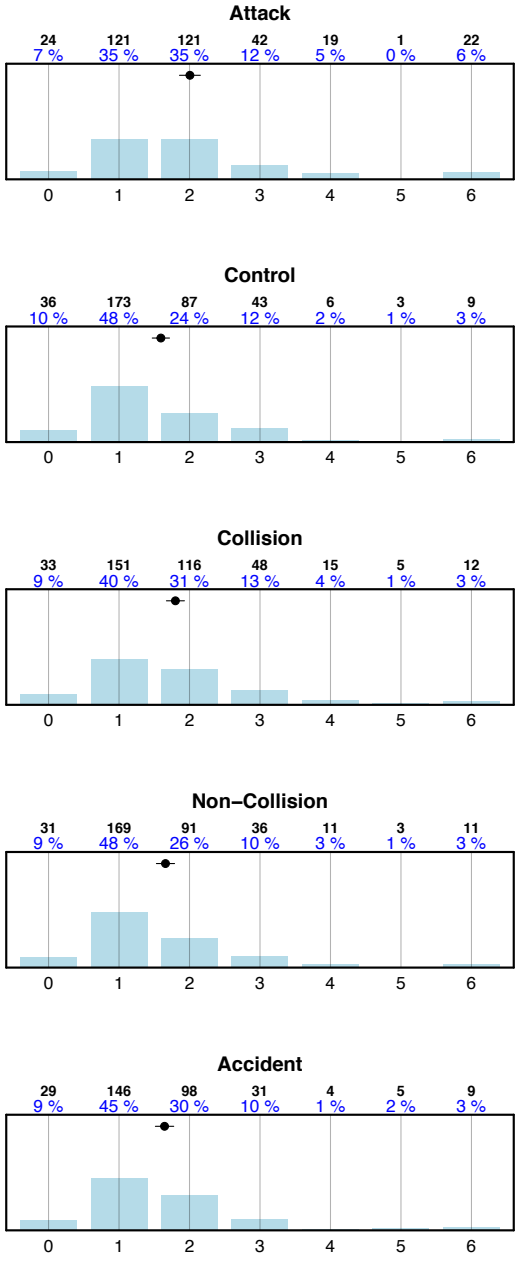

Figure 15: Distribution of Responses: Risk of War. For the **Risk of War** variable, 0 means the US should not assert its claims; 1 means the US should assert its claims only if can avoid any risk of war with China; 2 means the US should assert its claims only if it leads to at most a low risk of war with China; 3 means the US should assert its claims only if leads to at most a medium risk of war with China; 4 means the US should assert its claims even if it leads to a high risk of war with China; 5 means the US should assert its claims even if it leads to a very high risk of war with China; 6 means the US should assert its claims no matter the risk of war with China.

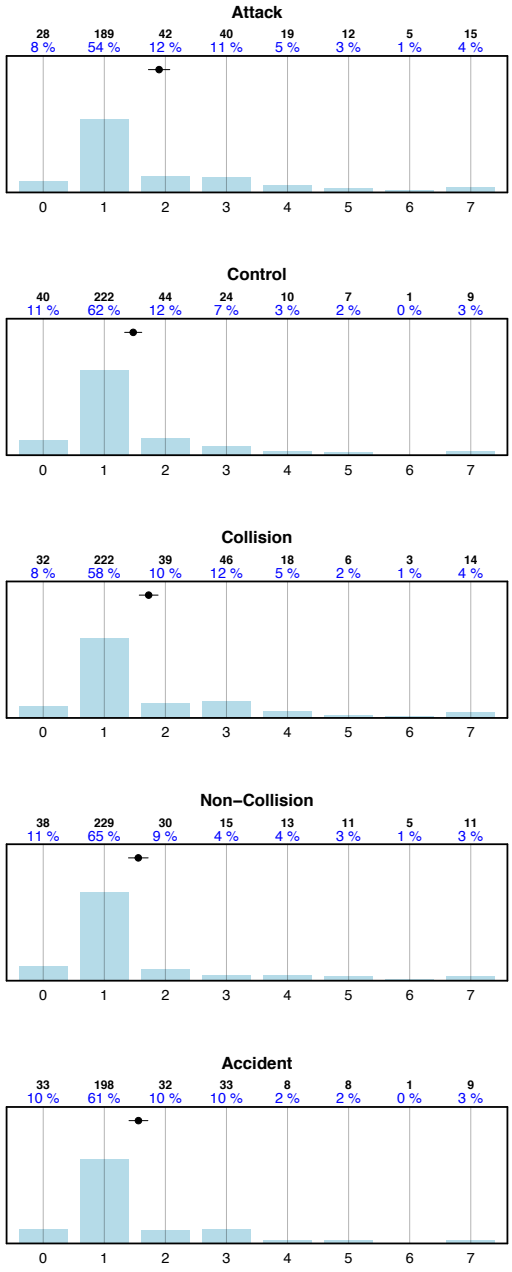

Figure 16: Distribution of Responses: Military Deaths. For the **Military Deaths** variable, 0 means the US should not assert its claims; 1 means the US should assert its claims only if can avoid US military fatalities; 2 means 3 military fatalities; 3 means 30 military fatalities; 4 means 300 military fatalities; 5 means 3,000 military fatalities; 6 means 30,000 military fatalities; 7 means the US should assert its claims no matter the number of military fatalities.

C.4.1 Reputation and Information Mediators

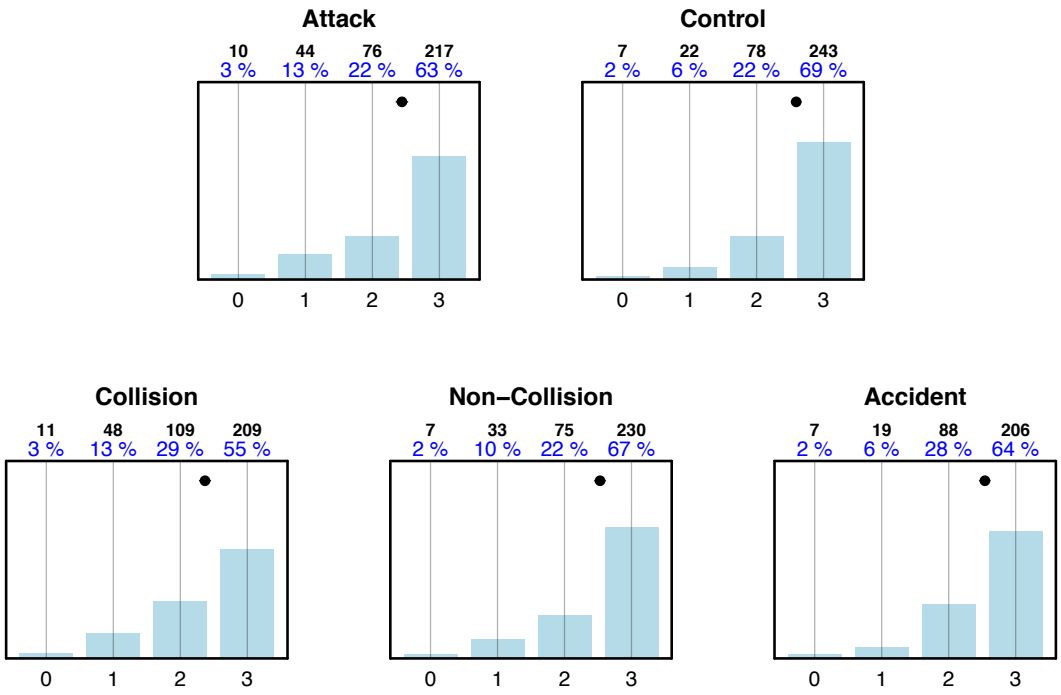

Figure 17: Distribution of Responses: Respondents' Considerations When Deciding to Maintain Claims – Avoid Risk of Conflict

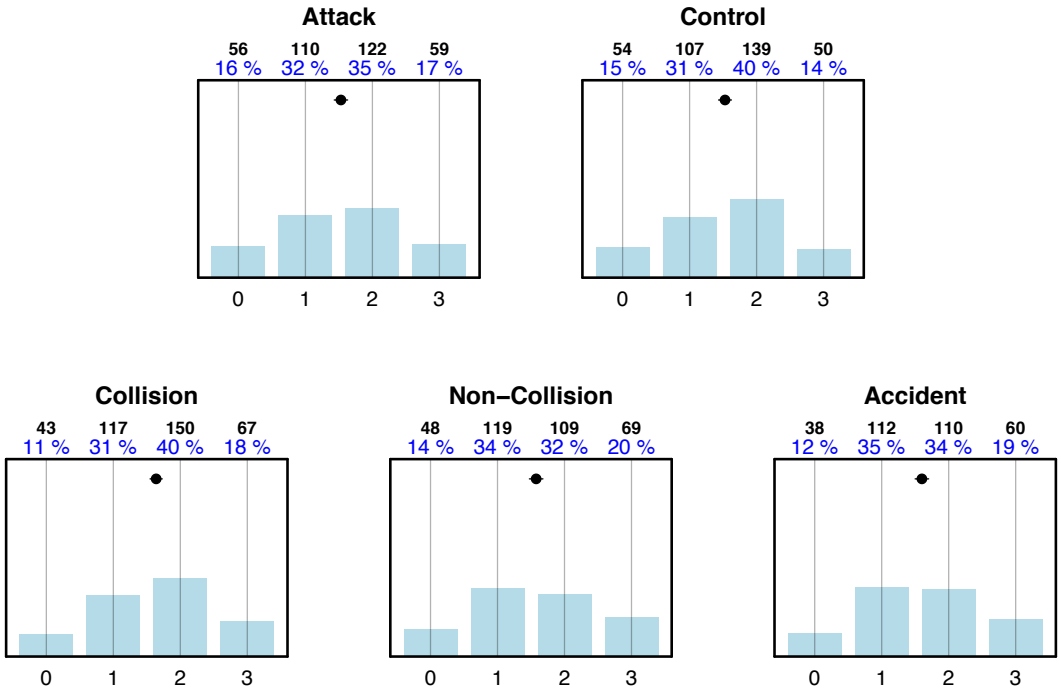

Figure 18: Distribution of Responses: Respondents' Considerations When Deciding to Maintain Claims – Preserve US Access to E. China Sea

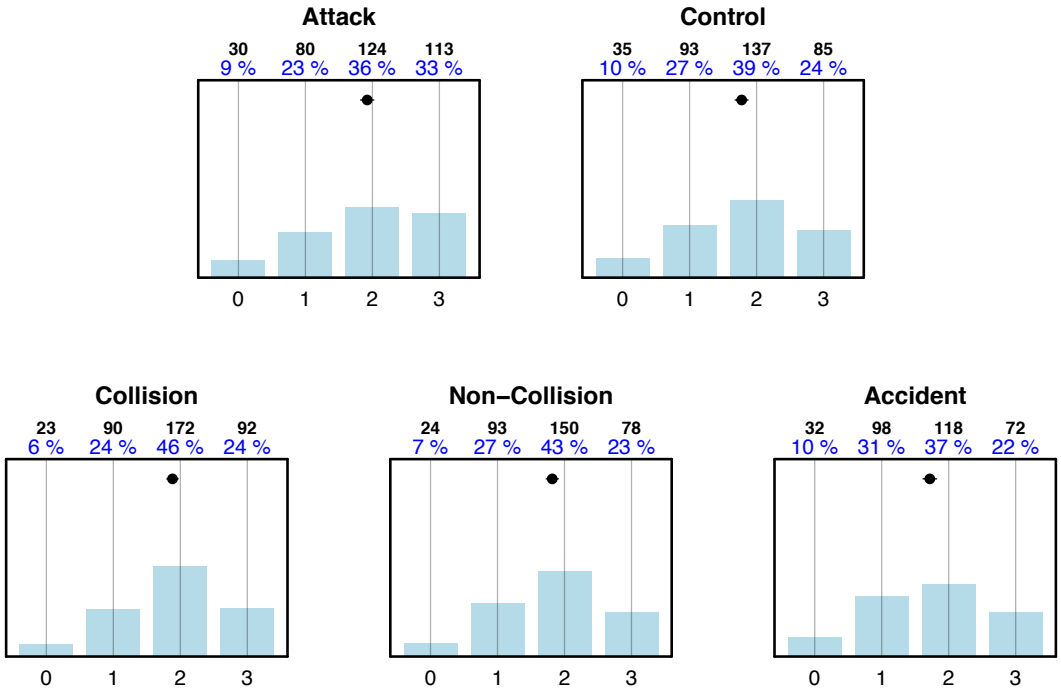

Figure 19: Distribution of Responses: Respondents' Considerations When Deciding to Maintain Claims – Resist China's Threat to US

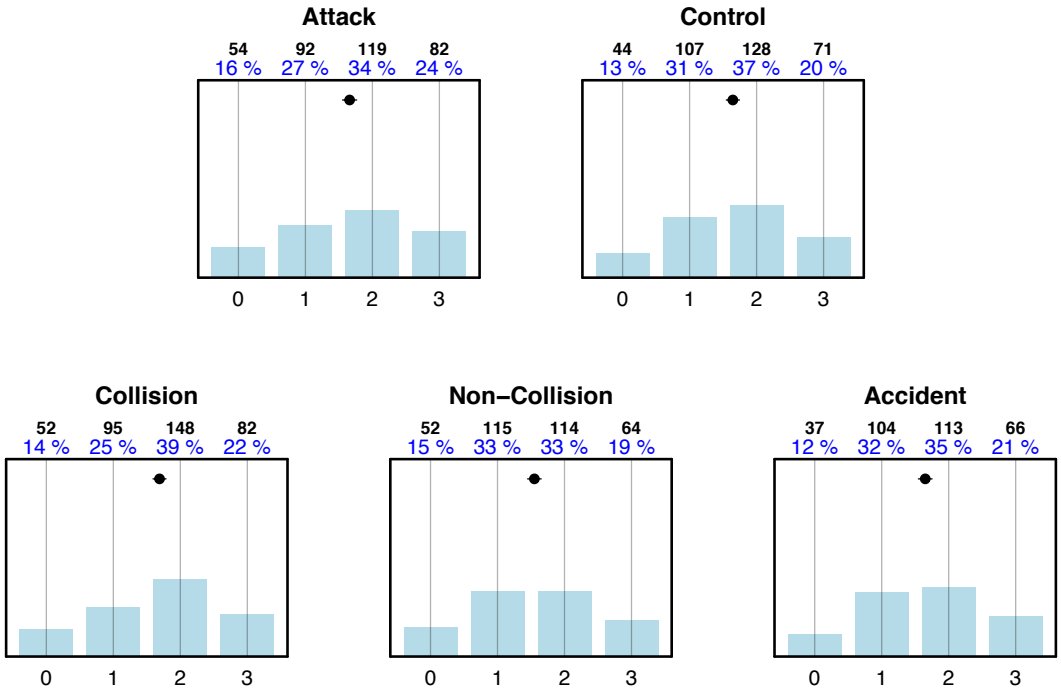

Figure 20: Distribution of Responses: Respondents' Considerations When Deciding to Maintain Claims – Maintaining US Reputation

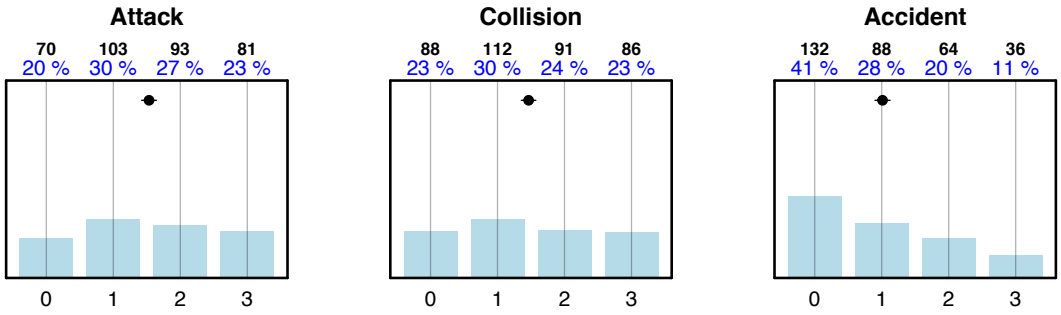

Figure 21: Distribution of Responses: Respondents' Considerations When Deciding to Maintain Claims – Honor US Pilot

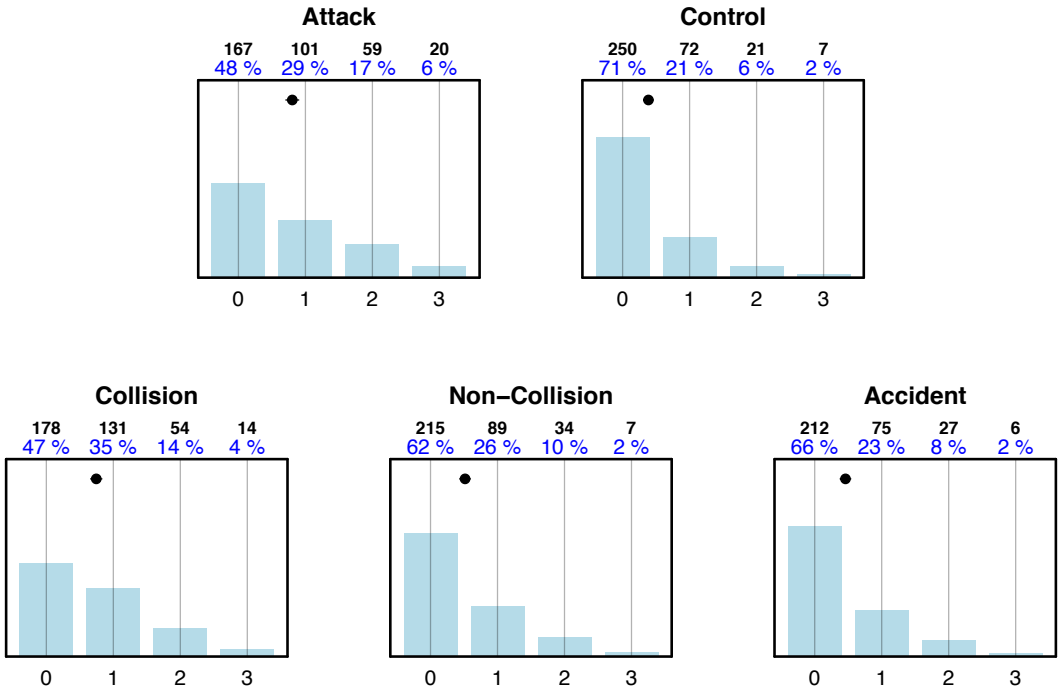

Figure 22: Distribution of Responses: Respondents' Considerations When Deciding to Maintain Claims – Punishing China

C.5 Subgroup Analysis

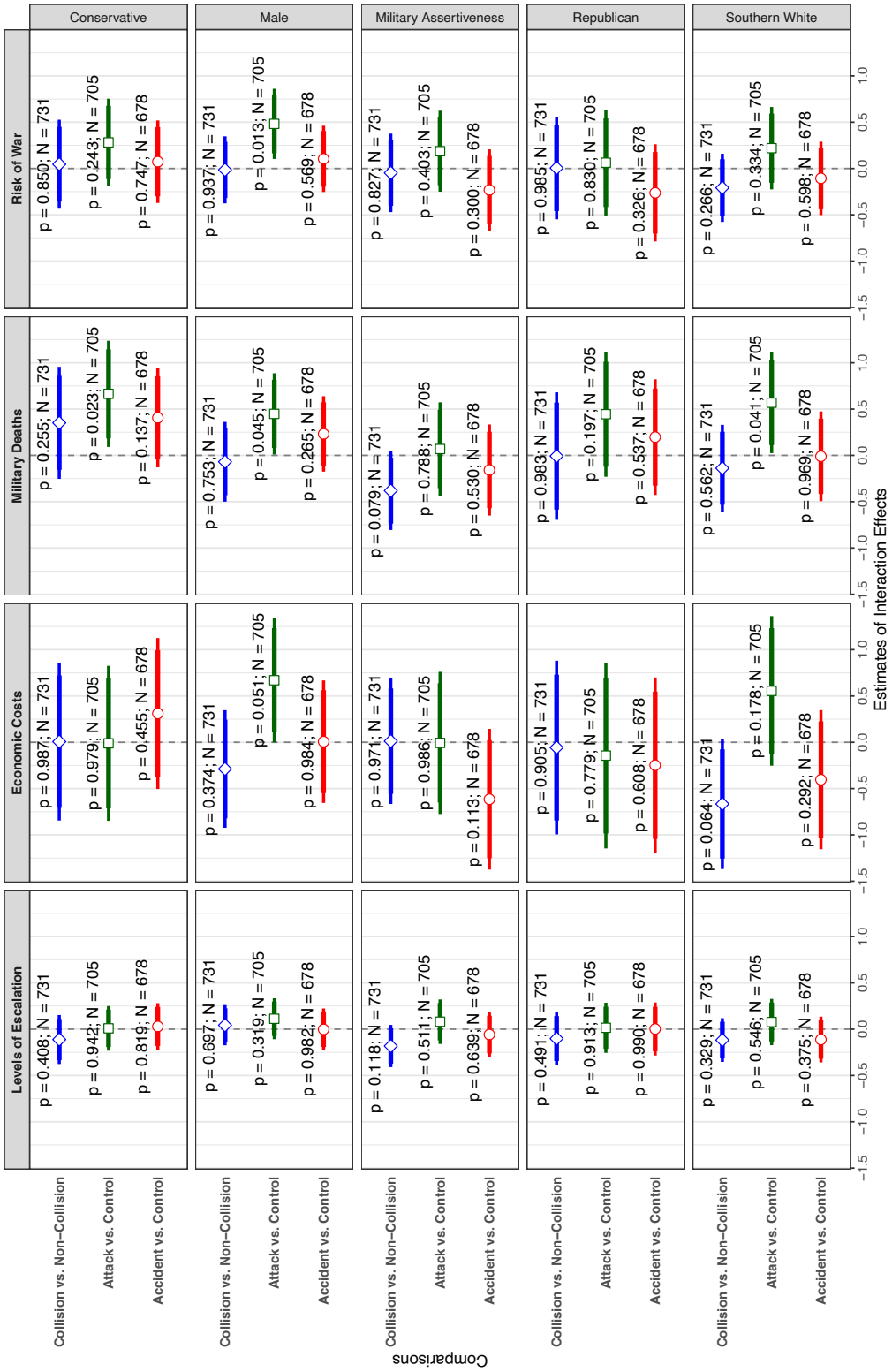

Figure 23: Subgroup Analysis of Main Outcome Effects. We perform subgroup analysis for males, Southern white, Republicans, Conservatives, and military assertiveness. (Southern states include South Carolina, Mississippi, Florida, Alabama, Georgia, Louisiana, Texas, Virginia, Arkansas, Tennessee, North Carolina, Arizona, Kentucky, New Mexico, Oklahoma, and West Virginia. Conservatives are those whose political ideology score is 5 or above on a 7-point scale. Military assertiveness is measured along a 5-point scale, such that higher values means greater military assertiveness.) For each outcome variable, we regress the responses on the treatment assignment, the subgroup variable, and the treatment assignment interacted with the subgroup variable. In the figure above, we present the estimated interaction effects for each main outcome and scenario comparison.

C.6 Descriptive Statistics

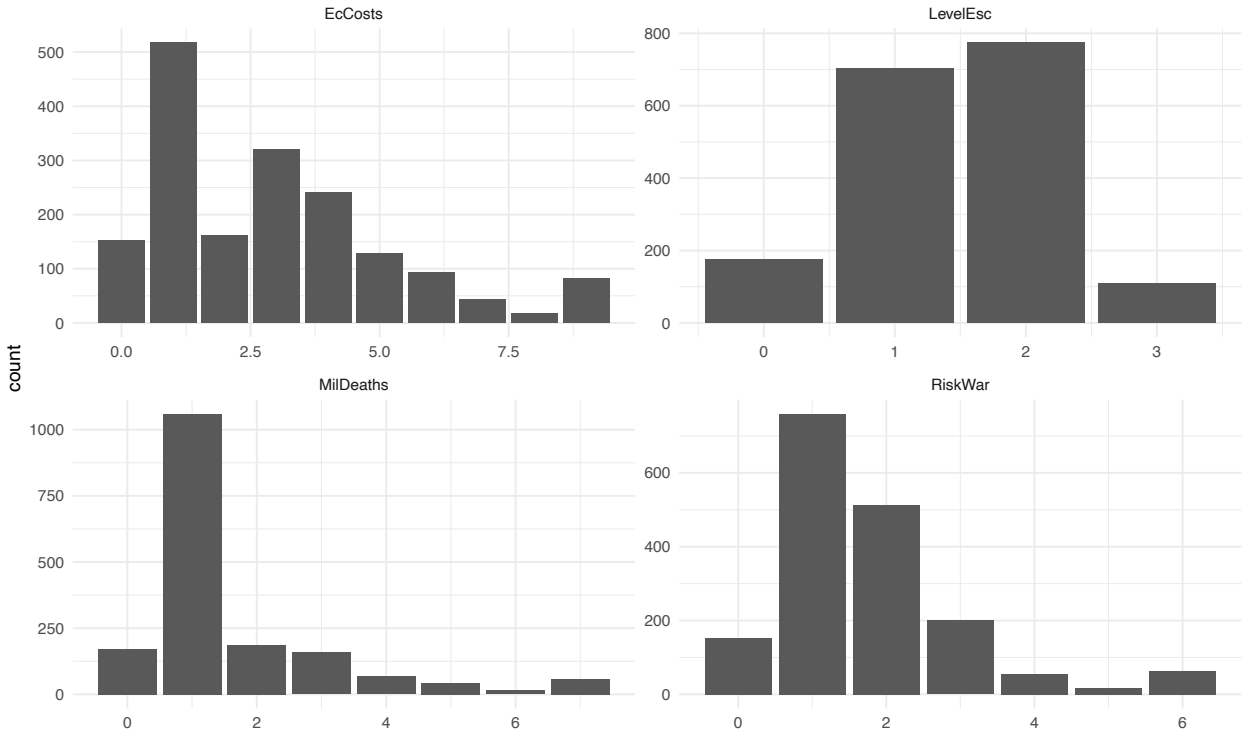

Figure 24: Distribution of Responses: Outcome measures

	n	min	max	mean	sd
LevelEsc	1761	0	3	1.46	0.76
EcCosts	1761	0	9	2.94	2.30
RiskWar	1761	0	6	1.74	1.27
MilDeaths	1761	0	7	1.64	1.51

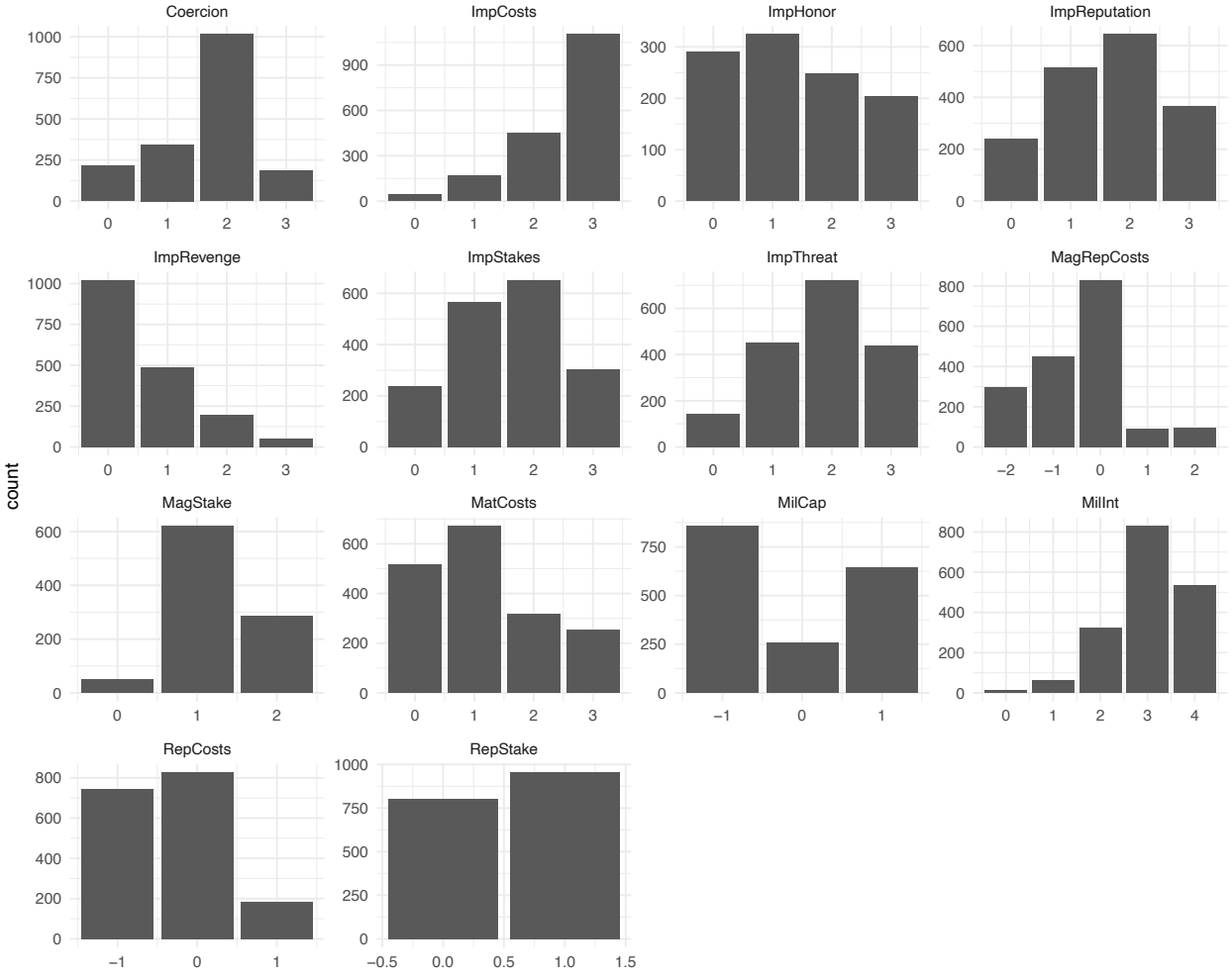

Figure 25: Distribution of Responses: Mediator measures

	n	min	max	mean	sd
RepStake	1761	0	1	0.54	0.50
MagStake	940	0	2	1.25	0.54
MatCosts	1761	0	3	1.18	1.01
Coercion	1742	0	3	1.66	0.83
MilCap	1761	-1	1	-0.12	0.92
MilInt	1761	0	4	3.03	0.83
RepCosts	1761	-1	1	-0.32	0.65
MagRepCosts	1761	-2	2	-0.43	1.00
ImpCosts	1739	0	3	2.49	0.77
ImpStakes	1739	0	3	1.58	0.93
ImpThreat	1739	0	3	1.83	0.90
ImpReputation	1739	0	3	1.64	0.96
ImpHonor	1044	0	3	1.35	1.08
ImpRevenge	1739	0	3	0.59	0.81

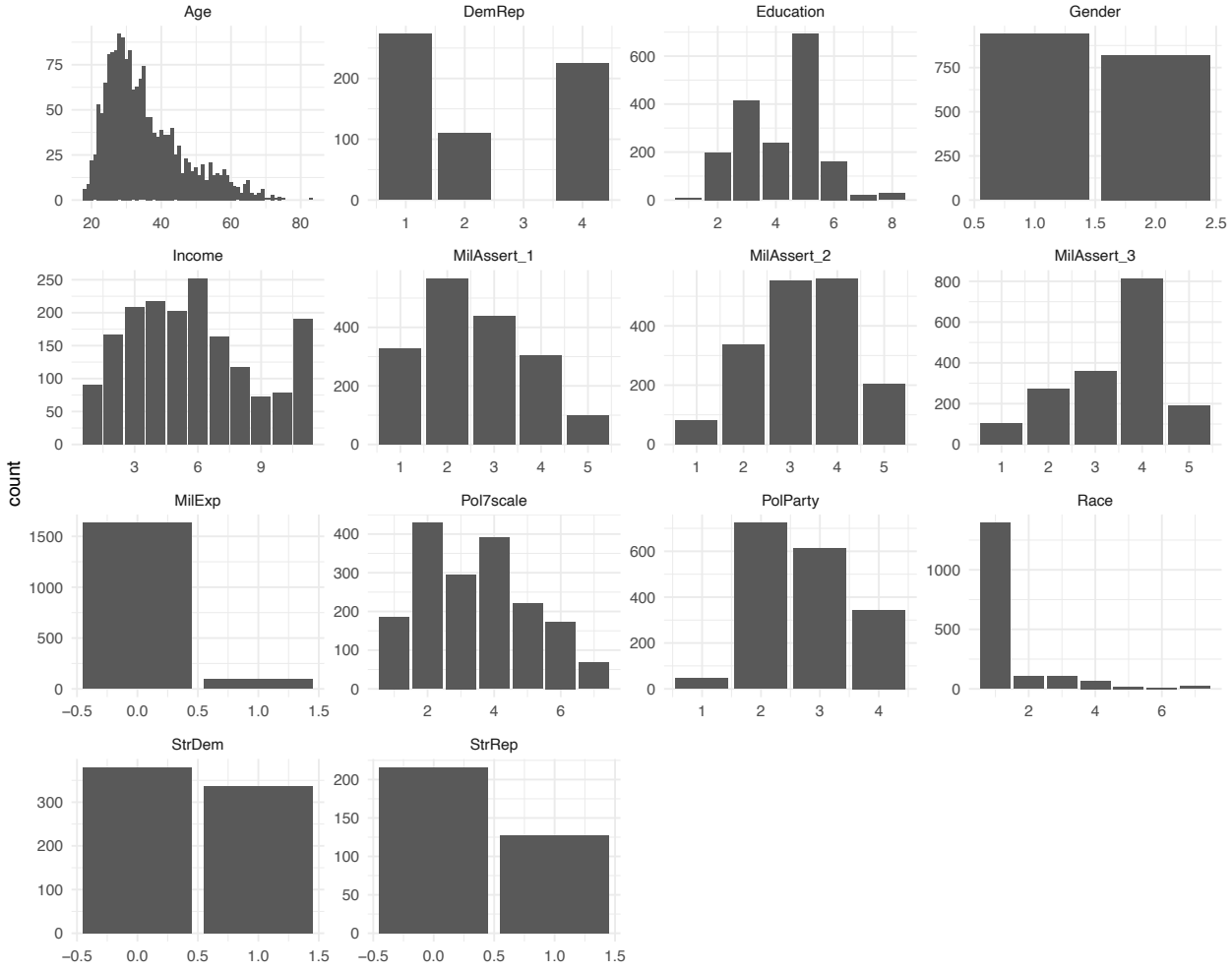

Figure 26: Distribution of Responses: Covariate measures

	n	min	max	mean	sd
Gender	1732	1	2	1.47	0.50
Age	1737	0	1982	36.55	48.08
Education	1732	1	8	4.21	1.35
Race	1723	1	7	1.45	1.11
Income	1690	1	11	5.71	2.99
Pol7scale	1723	1	8	3.55	1.80
PolParty	1678	1	4	2.72	0.82
StrDem	716	0	1	0.47	0.50
StrRep	343	0	1	0.37	0.48
DemRep	608	1	4	2.29	1.36
MilExp	1729	0	1	0.06	0.23
MilAssert_1	1735	1	5	2.58	1.15
MilAssert_2	1735	1	5	3.27	1.05
MilAssert_3	1735	1	5	3.41	1.06

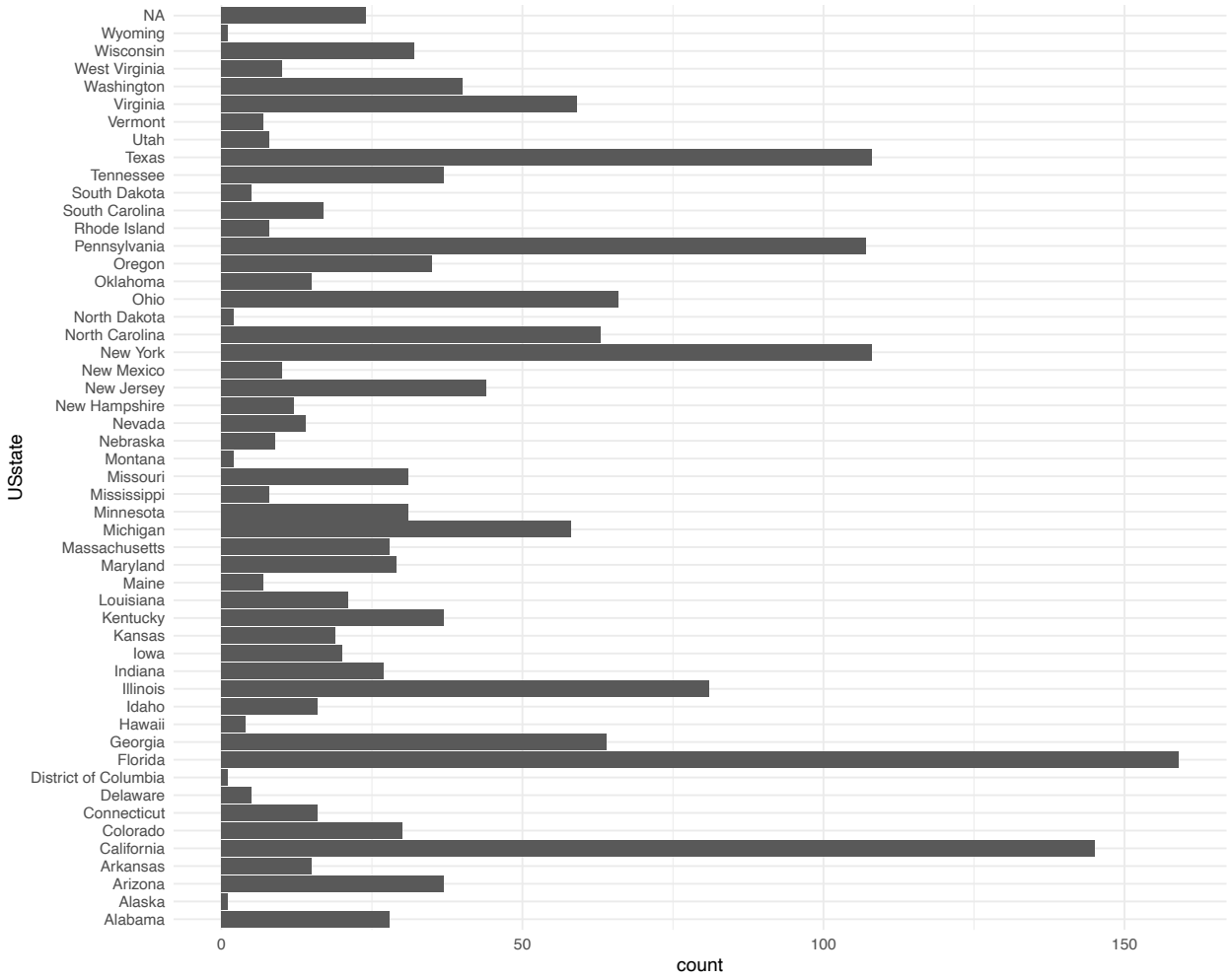

Figure 27: Distribution of Responses: US State

D Coercion

D.1 Brute Force and Schelling-Coercion

We define *coercion* as an attempt to get something from a target, done in a manner that makes the target of the attempt (weakly) worse off than absent the attempt. This definition is more expansive than typical definitions of coercion in IR, which follow Schelling in distinguishing "brute force" from coercion ((Ellsberg 1968, 3) similarly distinguishes the "ability to compel", or "naked power", from the ability to coerce). Brute force involves simply "taking what you want" (Schelling 1966, Loc 170) whereas most other forms of coercion involve getting the target to choose to comply. For clarity, we refer to definitions of coercion that exclude brute force as *Schelling-coercion*. While we think Schelling's distinction is a useful theoretical distinction, we think this distinction is not practically useful and has caused considerable confusion in the literature. We discuss several reasons why the Schelling definition is problematic.

Military action, even when it is intended as Schelling-coercion, tends to follow the idiom of the brute force pursuit of one's objective. This makes it practically difficult or impossible to separate Schelling-coercion from attempts at brute force. Consider the major disputes between the US and Soviet Union. In the Berlin Blockade, the Soviet's used brute force to block Western access to West Berlin. The US escalated by (brute force) flying in supplies using cargo planes. The natural response for the Soviets would have been to (brute force) shoot down these planes. Each step in this escalation followed the logic of each side using brute force to do precisely what it wanted to do, and yet the whole process is much better understood as a contest in risk taking: an instance of Schelling-coercion. Similarly, while the Cuban Missile Crisis is best understood as a contest in risk taking, each action followed the idiom of brute force. The Soviet's sent missiles and began deploying the missiles. The US imposed a blockade and considered air strikes against the missile sites. The Soviet's fired at the reconnaissance planes flying over Cuba; the US almost retaliated against the attacking SAM sites. Again, each action follows the brute force idiom of the pursuit of the state's objectives, and yet the whole process is better understood as a contest in risk taking (on this point, see Schelling 1966, §4). States usually confine themselves to attacking capabilities and imposing costs in a simple way in the specific issue dispute, rather than exercising the many other means of coercion at their disposal. The reason that escalation tends to follow this idiom is a separate issue; what matters is that coercion is usually expressed in the idiom of the brute force, direct pursuit of one's objectives on the issue in dispute. It is then not practical to distinguish brute force from Schelling coercion, since most Schelling coercion takes place through brute force actions.

This inextricable mixture of brute force and Schelling-coercion is apparent when we consider degradation attacks. The literature has noted that coercion often involves the degradation of the target's capabilities, such as by removing the leaders of terrorist organizations or air-strikes against military assets. Degradation attacks increase the costs of resistance, and reduce the expected utility from fighting. Degradation attacks can thus be thought of as Schelling-coercion because they change the utility calculation of the target, making the target more willing to comply. But degradation attacks also involve brute-force: they use brute force, and not persuasion, to change the target's situation. Degradation attacks thus embody both brute force and Schelling-coercion: they change the facts of the situation through "naked power", and these changed facts change the utility

calculation of the target to make acquiescence more appealing. At the limit, we can even define brute-force as a degradation attack that pushes the costs of resistance to infinity, or the expected utility of fighting to negative infinity. In this sense, brute force is simply a limiting case of degradation attacks. To summarize, degradation attacks involve a non-separable mixture of brute force and Schelling-coercion: they involve brute force actions that can Schelling-coerce.

We have argued that most examples of Schelling-coercion involve brute force. Similarly, most examples of brute force also involve Schelling-coercion. A classic example of brute force is the conquest of a territory through the forceful removal or destruction of a target army. And yet, except for the limiting case of complete physical domination and control, in practice brute force actions rely on Schelling-coercion to be successful. You may be able to seize a territory with force, but if the target refuses to acquiesce to this conquest you will have to repeatedly fight wars and a never-ending insurgency to hold on to the territory. Consider the control of inmates in a high-security prison. While, if necessary, the inmates can be compelled through brute force, in practice forcefully compelling every inmate to do everything would be excessively costly; instead Schelling-coercion is employed extensively.

Thus, Schelling-coercion almost always involves brute force, and brute force is almost always supplemented with Schelling-coercion. Pure brute force is only employed when there Schelling-coercion can not be used to supplement it, and Schelling-coercion follows the idiom of brute force. In practice the two are inextricably intertwined.

The term coercion, as generally understood outside of international relations scholarship, encompasses both Schelling-coercion and brute force. Were it not so, we would need another term as its substitute to refer to the large set of phenomena that involve some mixture of Schelling-coercion and brute force. We are not aware of such an alternative term. Accordingly, we recommend that IR scholars define coercion to include brute force, though recognizing that in practice most instances of coercion involve Schelling-coercion. Doing so will allow us to clarify conversations about whether air-strikes are instances of brute force or Schelling-coercion: in principle they are both; in practice some uses of air power will be employed for the purpose of altering the target's utility calculation, and some for the direct purpose of physically changing the facts on the ground.

An advantage of the more general definition of coercion is that it fits better with the logic of reputation and honor. An agent's reputation for resolve and honor is harmed by both Schelling-coercion and brute-force. This distinction is thus not useful for our theory, and we prefer to avoid it by referring to both mechanisms as coercion.

D.2 Does Coercion Require Success?

A related conceptual and definitional issue is whether the term coercion, or here "Schelling-coercion", implies *successfully* altering the other target's behavior in the way the coercer wants.[6] Some scholars use coercion in this way, and some usage of cognate terms is consistent with this. For example, Nozick's influential paper defines "coerced" as requiring that the person changed their behavior because of the threat: that the threat succeeded.

The appropriate definition depends on a scholar's goals. In our case, for theoretical reasons our recommended usage and definition does not depend on whether the coercive act succeeds. If coercion only exists when the attempt to coerce succeeds, then one can't study whether coercion

[6]Thanks to a reviewer for raising the need to clarify this point.

can backfire. Coercion, by this alternative definition, always succeeds. One could replace the term coercion with "attempted coercion", and ask when "attempted coercion" succeeds or backfires; however, this is cumbersome and not necessary, and also not consistent with broad usage of the term. For example, a victim objecting that "this is coercion", or a policy analyst noting that a "policy of coercion was useless", are using the term in the sense of attempting to coerce.

D.3 Coercion, Offers, Information

A final important distinction is between coercion (sometimes called extortion or blackmail), and other activities that may at times look similar because they involve trying to get an actor to change their behavior, such as offering positive inducements ("exchange") or providing information ("explanation"). See (Oye 1993, Chapter 3) for a helpful discussion of these three. The distinction between these is often difficult to establish in practice, in part because actors try to obfuscate them. We predict, as have others, that it is coercion that produces a provocative effect, and not offers or information. This is consistent with the logic of reputation and honor, since it is only coercion that one should seek to systematically deter.

D.4 Should "Coercion" Imply Success?

Coercion has also been widely discussed in ethics (Anderson 2017); with classical political theorists employing a behavioral conceptualization closer to ours, and Nozick (1969) and more recent philosophical works employing a conceptualisation closer to Schelling's (requiring that coercion operate only through incentives). Nozick, in defining what it means to be coerced, also required that the coercive act succeed; such a "success condition" would not be appropriate for our inquiry, which investigates when coercion does not succeed.