Coercion and the Credibility of Assurances

Matthew D. Cebul, United States Institute of Peace
Allan Dafoe, University of Oxford
Nuno P. Monteiro, Yale University

What makes coercion succeed? For most international relations scholars, the answer is credible threats. Yet scholars have neglected a second key component of successful coercion: credible assurances. This article makes two contributions to our understanding of coercion and credible assurances. First, we offer a theoretical framework exploring the causes and consequences of assurance credibility. In order to coerce the target, a challenger must issue both credible threats that defiance will be met with punishment, and credible assurances that compliance will be met with restraint. In turn, the credibility of assurances is determined by power and a reputation for restraint. Whereas greater power boosts credible threats, it undermines credible assurances. Therefore, powerful states must cultivate a reputation for restraint in order to issue credible assurances. Second, we provide empirical support for our claims through a nationally representative, scenario-based survey experiment that explores how US citizens respond to a hypothetical coercive dispute with China.

The United States often fails at coercion. In 1999, Serbian President Milošević ignored President Clinton's demands for Serbian troops to leave Kosovo, leading to a NATO bombing campaign. Iraqi President Hussein rebuffed American demands in both 1990–91 and 2002–3, prompting two Gulf Wars. And decades of US pressure has failed to convince North Korea to abandon nuclear weapons development. These examples are part of a broader trend. Examining 15 prominent cases of US coercive diplomacy, Art and Cronin (2003) find no complete success stories and observe more failures than even qualified successes.

This meager record of US coercive success defies the dominant understanding of coercion in international relations. Most scholarship identifies threat credibility as the key to successful coercion, highlighting two determinants of credible threats: power and a reputation for resolve.[1] The United States has been the most powerful state in the world throughout the postwar era, especially since the end of the Cold War. Its unparalleled military gives Washington unprecedented freedom of action on the world stage. Its vigorous foreign policy has established a reputation for resolve. American threats of punishment are therefore highly credible. What, then, explains the United States' failure to coerce adversaries?

This article seeks to reconcile our theoretical understanding of coercion—powerful states with a reputation for resolve issue credible threats of punishment and therefore succeed at coercion—with its empirical reality: powerful, resolved states are often thwarted in their coercive efforts. To do so, we shift the focus from threats to assurances against future punishment, offering a theory of coercion that gives equal weight to each. While the sources and role of threat credibility in successful coercion are the object of abundant scholarship,

Matthew D. Cebul (mcebul@usip.org) is a research officer at the United States Institute of Peace, Washington, DC 20037. Allan Dafoe (allan.dafoe@politics.ox.ac.uk) is an associate professor of international politics at the University of Oxford, Oxford, UK OX1 1PT.

On behalf of his many students and colleagues, we dedicate this article to Nuno Monteiro, who possessed brilliance and kindness in equal measure, and who guided us in both scholarship and life; and to his surviving family, who shared his precious time with us. Together, may we carry Nuno's legacy as he so often encouraged us—onward. Supplementary material for this article can be found in the online appendix. Replication files, and links to our preregistration of our experimental design, are available in the *JOP* Data Archive on Dataverse (https://dataverse.harvard.edu/dataverse/jop). Online survey research was deemed exempt by the Yale Institutional Review Board, Human Subjects Committee. Support for this research was generously provided by the Yale MacMillan Center.

1. We use the term "resolve" to describe the challenger's willingness to follow through on its threats and the target's willingness to resist demands, following Dafoe and Caughey's (2016, 6) definition of "resolve" as an "actor's willingness to stand firm in a particular dispute." This broad conception builds from earlier definitions based on the relative attractiveness of war versus conceding, whereby resolve is "the maximum risk of bargaining breakdown . . . that the party is willing to accept in standing firm" (Snyder and Diesing 1977, 191; see also Morrow 1989, 941). In this sense, resolve is usually understood as a function of various factors, including the balance of capabilities, the value of the object at stake, and the costs of war. Others describe resolve as an inherent dispositional or psychological trait (Kertzer 2016), be it a "willingness to take risks" (Sartori 2005, 7) or willingness to endure costs, the latter being especially prominent in studies of casualty aversion and war in democracies (Gelpi, Feaver, and Reifler 2006; Stam 1996).

The Journal of Politics, volume 83, number 3. Published online May 20, 2021. https://doi.org/10.1086/711132

assurance credibility has only recently begun to receive the attention it deserves (Sechser 2010, 2016), and much work remains to be done.

To succeed, coercive attempts—be they positive coercion (compellence) or negative coercion (deterrence)—must include two aspects: credible threats of punishment if the target ignores the challenger's demand and credible assurances of restraint if the target acquiesces. When these two aspects are present, a coercive attempt is balanced and more likely to succeed. The absence of either undermines this balance and lessens the odds of coercive success. As the existing literature emphasizes, absent credible threats the target is more likely to ignore the challenger's demand because it expects that defiance will not be punished. As this article stresses, absent credible assurances the target is also more likely to defy the challenger's demand, in this case because it expects punishment even if it complies.

Making assurances credible while also issuing credible threats is not easy, however. Relative power, a key determinant of both threat and assurance credibility, pulls threats and assurances in opposing directions: power boosts the credibility of threats but undermines that of assurances. To issue credible threats and assurances simultaneously, a powerful challenger must resort to the second key determinant of the credibility of assurances: a reputation for restraint in the face of compliance with its past coercive demands. By highlighting these dynamics, our theory contributes to the broader debate on the relative importance of power and reputation for international crisis bargaining.

We support our claims with data from a scenario-based survey experiment that gauges US citizens' reaction to a hypothetical confrontation with China, manipulating China's relative power and its past history of restraint. We find that (i) US citizens are more likely to abide China's demands when Beijing issues both credible threats and credible assurances, that (ii) increasing China's power boosts the credibility of its threats but undermines that of its assurances, and that (iii) having a reputation for restraint improves the credibility of China's assurances.

COERCION, THREATS, AND ASSURANCES

Thomas Schelling clearly noted the centrality of credible assurances to successful coercion. As he states: "Any coercive threat requires corresponding assurances; the object of a threat is to give somebody a choice. To say, 'One more step and I shoot,' can be a deterrent threat only if accompanied by the implicit assurance, 'And if you stop I won't.' Giving notice of *unconditional* intent to shoot gives him no choice" (Schelling 1966, 74, Schelling's emphasis). Notably, Schelling (1966, 74–75) goes on to caution against overemphasizing threats at the expense of assurances; although the relationship between threats and coercion is intuitive, assurances are also vital to coercive success.

Scholars of coercion, however, have largely neglected Schelling's insights about assurances, instead devoting vastly more attention to the credibility of threats. Scholars have debated at length the sources of threat credibility, gravitating toward two key determinants: relative power (Hopf 1994; Mercer 1996; Press 2004/5, 2005) and a reputation for resolve based on past behavior (Clare and Danilovic 2010; Crescenzi 2007; Gibler 2008; Guisinger and Smith 2002; Harvey and Mitton 2016; Huth 1997; Sartori 2005; Weisiger and Yarhi-Milo 2015). Beyond these two central factors, others have emphasized the importance of regime type (Fearon 1994, 1997; Gelpi and Griesdorf 2001; Partell and Palmer 1999; Weeks 2008). While these scholars disagree on what makes threats credible, all accept the fundamental premise that threat credibility is the linchpin of successful coercion.

This emphasis on threats is a legacy of Cold War politics; in the context of mutually assured nuclear destruction between the two superpowers, assurances that compliance with coercive demands would be met with restraint, even if implicit, seemed eminently credible. Along these lines, Schelling observed that the credibility of assurances was more evident in the context of nuclear deterrence, as deterrent assurances are naturally reinforced in advance by the deterring state's status quo inaction.[2]

The demise of the Soviet Union almost three decades ago, however, radically transformed the strategic background to US coercive efforts. Since then, the United States has enjoyed a preponderance of power, and its coercive ambitions have shifted from deterring Soviet aggression to compelling behavioral changes in recalcitrant states. In this context, implicit US assurances that compliance with coercive demands will be met with restraint can no longer be taken for granted.

Despite this sea change, most contemporary studies of coercion persist in emphasizing the role of credible threats. Yet these threat-centric theories struggle to explain the puzzling fact that the United States is frequently unable to coerce its adversaries. Excluding the eight other nuclear-weapons states and their handful of protégés, the United States has the military capacity to impose great costs practically anywhere

2. As Schelling (1966, 74–75) writes, "The assurances that accompany a compellent action—move back a mile and I won't shoot (otherwise I shall) and I won't try again for the second mile—are harder to demonstrate in advance, unless it be through a long past record of abiding by one's own verbal assurances." These insights presage our thinking about the value of a reputation for restraint.

in the world with relative impunity, and indeed it has frequently used military force to achieve its objectives (Monteiro 2014). Therefore, both the power and reputation schools of threat credibility would predict that American threats are highly credible, perhaps more so than at any other time in history. Yet states regularly resist American coercion.

In turn, coercion scholars have generated several explanations for this disjuncture. Some argue that US threats are not credible precisely because Washington enjoys a preponderance of power, which makes threats cheap to carry out and therefore undermines their signaling value (Chamberlain 2016). Others argue that, for a variety of reasons, US adversaries often misperceive US capabilities or its reputation for resolve (Jervis 2003). Although valuable contributions, these arguments start from the assumption that failures in US coercion stem from a real or perceived lack of credible threats.

Schelling's insights about the need for corresponding assurances suggests an alternative possibility: American coercion may fail despite credible threats because it is not complemented by credible assurances. In fact, critical analyses of several prominent US coercive efforts highlight the importance of credible assurances. For example, Christensen (1992, 136) argues that the United States failed to deter China from entering the Korean War because Mao disbelieved the credibility of implied US assurances against future conflict. As he writes, "America's crossing of the parallel [on October 7, 1950] convinced Mao of America's aggressive intent and led him to dismiss subsequent American promises of restraint. . . . The expansion of American military operations in Korea convinced Mao that a war between American and Chinese forces was inevitable, regardless of the details of subsequent American coercive diplomacy" (Christensen 1992, 136). In more recent history, the Bush administration failed to recognize that Saddam Hussein had little reason to comply with American demands in 2002–3 if he thought that Washington would pursue regime change even if he cooperated. As Jervis (2003, 325) writes, "the administration seems to have trouble with Schelling's basic point that if the other side is to be influenced, threats to act if the other refuses to comply with demands must be paired with promises not to take the action if the other does cooperate. . . . American threats have been undercut by the refusal to promise that Saddam could stay in power if he gives up the forbidden weapons."[3] On a positive note, US assurances appear to have helped facilitate successful US coercion of Libya in 2003, as Muammar al-Gadhafi repeatedly sought and received assurances that the United States would not press for regime change in Libya before dismantling his nascent nuclear program (Jakobsen 2012; Jentleson and Whytock 2005/6), though arguably the United States later reneged on these assurances when it militarily intervened against the regime in 2011.

Despite these historical examples, the literature on coercive assurances remains limited.[4] Some scholars, following Schelling, note the basic need for symmetry between credible threats and assurances but do not investigate the determinants of assurance credibility (Myerson 2006; Stein 1991).[5] Only a few works directly analyze assurance credibility in coercive diplomacy; we briefly review them below.

To start, Abrahms (2013) argues that terrorists' efforts to coerce states are characterized by a "credibility paradox": "the very escalatory acts that add credibility to a challenger's threat can subtract credibility from his promise [of restraint]" (Abrahms 2013, 660). In his view, this paradox is the product of a cognitive heuristic that leads humans to "confound the extreme means of the challenger with his presumed ends" (661). Because terrorists often employ extreme violence, states doubt the credibility of their assurances of future restraint and thus refuse to cooperate despite credible threats of terrorism. While Abrahms's framework can be generalized beyond terrorism, it presents two limitations. First, it does not permit variation in the credibility of assurances relative to that of

3. See also Hiroshima (2015). There is also some evidence that Saddam irrationally believed that US aggression would be limited and restrained by other global powers—see Lake (2010). Nevertheless, it is clear that Saddam and his cabinet were deeply concerned about the United States' desire for regime change throughout the 1990s. For example, Hiroshima (2015, 38–39) cites foreign minister Tariq Aziz's suspicion that UNSCOM inspection demands in the aftermath of the first Gulf War presaged further US hostility; as Aziz states, "even if we implement and moved from 20% to 30% to 40% to 50% to 60% to 70% in implementation and there is no advancement from your [UNSCOM] side as if we did not implement anything . . . you hit us in the past and now you are threatening us that you could attack us. So we do not have any guarantee that you will not attack us again either if we implement or not implement because the motivations are political."

4. For unpublished work, see Monteiro (2009) and Pauly (2019). In contrast, there is a vast literature on "strategic reassurance," efforts by one state to impress upon its adversaries that its intentions are benign, thereby mitigating the mutual fear produced by the security dilemma (Knopf 2012a, 2012b; Kydd 2000; Midford 2010; Stein 1991). There is also a growing literature on "allied reassurance," efforts by one state to impress on its protégés its resolve to protect them from common adversaries, particularly in the context of nuclear alliances (Debs and Monteiro 2016; Fuhrmann and Sechser 2014; Lanoszka 2018).

5. The most recent of these is Wolford (2019, chap. 13), which presents a fascinating case study of Germany's decision to renew submarine warfare against US vessels in early 1917, ultimately dragging the United States into World War I. Wolford persuasively argues that Germany did so because German leaders came to believe that US intervention was largely inevitable, given that the United States was strongly incentivized to protect its investments in the Entente powers (who were heavily indebted to the United States), though he does not further elaborate on the sources of credible assurances of restraint.

threats—escalatory acts that make threats credible necessarily undermine assurance credibility. Second, it does not ground assurances within a baseline rationalist framework of coercion that does not rely on cognitive or psychological factors, such as the one we develop below.

For their part, Kydd and McManus (2015) develop a game-theoretic model of coercive assurances, finding that assurances can help resolve crises by deflating the challenger's minimum acceptable share of a disputed good. While this model integrates assurances into a rationalist framework, it does so by assuming that leaders pay a domestic audience cost for failing to uphold public assurances.[6] Apart from concerns that audience costs may be impotent signaling tools (Downes and Sechser 2012; Snyder and Borghard 2011), they are at best an indirect mechanism connecting assurance credibility to coercive success. Below, we present a general rationalist framework that more directly captures the role and sources of assurance credibility in coercion.

Last, Sechser (2010, 2016) most closely informs our work. Sechser (2010) argues that strong challengers often fail because weak targets resist coercion in order to build a reputation for toughness that could deter further demands. In turn, Sechser (2016) demonstrates that factors that increase the likelihood of future demands—including geographic proximity, the challenger's past history of aggression toward the target, and the challenger's ability to project power—increase the odds that the target resists the challenger's demands. Together, these studies helpfully illustrate that the shadow of future demands, and of assurances against them, looms large over coercive efforts. As Sechser (2010) writes, "even when a challenger's threats are completely credible, the balance of capabilities is publicly known, and settlements are enforceable, fears about a challenger's future intentions can motivate rational targets to fight losing wars to deter future aggression" (629).

We build upon, and depart from, Sechser's work on both theoretical and empirical fronts. First, we broaden our analysis of the determinants of assurance credibility, considering both relative power and the challenger's reputation for restraint. In doing so, we clarify the conditions under which strong challengers may nevertheless overcome "Goliath's Curse" to issue credible assurances. Second, while Sechser emphasizes the target state's desire to foster a reputation for toughness, we contend that the target's choice to reject demands can be explained more parsimoniously: the target rejects demands not because it hopes that a reputation for toughness will deter future demands, but because it deems that the challenger will pursue intolerable demands no matter the target's behavior. Because future aggression is hardly avoidable, the target gains nothing from acquiescence and therefore resists coercion to preserve whatever utility it accrues from the object under dispute.

Finally, we complement Sechser's (2016) cross-national observational findings with experimental evidence. The experimental approach advances the study of assurances in several ways. First, a controlled experimental environment permits the standardization of background factors, such as the nature of the demand or precrisis diplomacy, as well as a more controlled manipulation of causal factors of interest. This enables us to isolate more cleanly the hypothesized effects of power and reputation for restraint on coercive success. Second, large-*N* cross-national data may not speak to the mechanisms underlying observed relationships. In this case, while Sechser (2016) attributes (correctly, we think) the negative correlations between his independent variables and coercive success to the target's fear of future demands, he acknowledges that his data cannot directly prove that assurance credibility drives these effects (337). This is an important limitation, especially given that some of the factors he identifies are also plausibly linked to the credibility of threats, which could encourage the target to comply. In what follows, we measure assurance and threat credibility directly, establishing empirical support for the theoretical microfoundations that link assurance credibility to its determinants and, ultimately, to the challenger's resolve.

Wrapping up, as Knopf (2012b, 378) asserts, "security assurances are perhaps the least studied of all influence strategies that can be utilized by states." Consequently, the dominant framework through which scholars understand coercive diplomacy is imbalanced. Below, we address this problem by introducing a rational framework highlighting the role played by credible assurances in coercion, and the factors determining their credibility.

A THEORY OF CREDIBLE COERCIVE ASSURANCES

Successful coercion depends on the credibility of both threats and assurances. Without a credible threat of punishment if the target refuses to yield, the target will see little cost to defiance; without a credible assurance of restraint if the target does yield, the target will expect punishment to be inevitable, giving it little incentive to comply. We call this need for both threats and assurances the *balance of coercion*. Unbalanced coercive efforts—those from which either credible threats or credible assurances are missing—are more likely to fail.

It follows that when deciding whether to resist or comply with the challenger's demand, rational targets of coercion

6. For a supporting argument, see Levy et al. (2015), who find evidence that state leaders suffer audience costs when they "back into" conflicts after making public promises of restraint.

consider two questions. First, what is the likelihood that the challenger will follow through on its threat if I resist—that is, how resolved is the challenger? Second, what is the likelihood that the challenger will refrain from further demands if I comply—that is, how restrained is the challenger?

This section addresses the latter question by explicating the determinants of assurance credibility, emphasizing the importance of relative power and a reputation for restraint. Our theory of assurance credibility mirrors the ongoing debate on the role of power and a reputation for resolve as sources of threat credibility. At the same time, we complicate the current understanding of the role of power in shaping the odds of successful coercion.

Relative power

The role of relative power in coercive diplomacy is often misunderstood. Prior research correctly observes that relative power boosts the credibility of threats as, being more likely to prevail, powerful states are more willing to use force. Still, existing scholarship neglects the fact that relative power has the opposite effect on the credibility of assurances. Because a challenger with greater relative power enjoys a higher probability of victory in war, it is more willing to threaten conflict in the future even if the target complies with its present demand. Therefore, a powerful challenger will be more likely to issue subsequent demands. As a result, the more powerful a challenger, the greater the odds that the target will consider the current demand to be merely a prelude to future additional requests (Sechser 2010).[7]

Thus, relative power increases the credibility of the challenger's threats, while simultaneously decreasing that of its assurances. Theoretically, this means that the effect of power on the odds of coercive success is indeterminate: while some degree of power is necessary to generate credible threats (pushing the target to comply), excessive power can undermine the credibility of assurances (pushing the target to resist). Our theory is agnostic on the relative magnitude of these effects—increasing relative power may magnify threat credibility by more than it decreases assurance credibility in some cases, but do the reverse in others. It is possible that factors such as regime type, force posture, public statements, alliances, norms, and international institutions will condition the relative effect of an increase in the challenger's power on the credibility of threats and assurances. This means that the exact threshold beyond which power becomes more harmful than helpful to any particular coercive effort is difficult to describe ex ante.

We are thus left with the following hypotheses:

H1. All else equal, threat credibility is increasing in the challenger's relative power.

H2. All else equal, assurance credibility is decreasing in the challenger's relative power.

Reputation for restraint

This discussion suggests that strong challengers must turn to other factors to compensate for the depressing effect of relative power on assurance credibility. One such factor is a reputation for restraint. Existing scholarship has argued that targets use challengers' past behavior to gauge their likely behavior in current crises. Yet as noted above, these arguments have focused almost exclusively on threat credibility. We argue that past behavior also influences the credibility of assurances. A challenger that consistently stands by the agreements it strikes as the result of coercive efforts—and that forgoes making additional demands of compliant targets—will develop a reputation for restraint. In turn, this reputation boosts the credibility of the challenger's assurances against future demands.[8] In contrast, a target is unlikely to place much faith in assurances from a challenger that routinely oversteps its agreements and attempts to exploit opponents even when they fully comply.[9]

What is the form and function of reputations in international politics?[10] While scholars often speak of reputations as belonging to states, they may also adhere to leaders. Renshon, Dafoe, and Huth (2018) contend that reputations are more likely to adhere to leaders when leaders exert outsized influence over the relevant policy domain, though they also find

7. This discussion of power speaks to the rationalist literature on the "commitment problem" as a cause of war (Debs and Monteiro 2014; Fearon 1995; Powell 2006). In this tradition, commitment problems arise when proposed deals would produce large shifts in the balance of power and are therefore unenforceable. So defined, commitment problems are especially stark manifestations of the more general problem of assurance credibility in coercion, which obtains even if the issue at stake does not significantly alter the balance of power (as may be the case when the challenger is already especially powerful).

8. Mercer (1996) argues that favorable state behavior is often credited to situational attributes and that this context dependence precludes the formation of informative reputations. We agree that a state's past behavior may be interpreted differently depending on the observer's perspective, but believe that coercive situations are not so unique that states are unable to draw inferences from past behavior (see Crescenzi 2018; Harvey and Mitton 2016). Our results below provide experimental evidence that past behavior predictably influences respondents' assessment of a challenger's future intentions.

9. As we mention above, other factors such as regime type, force posture, public statements, alliances, norms, and international institutions may well condition a challenger's ability to issue credible assurances. Whether they do is a question we reserve for future research.

10. For a review of these topics, see Dafoe et al. (2014).

that reputations can transcend particular administrations. One example of a leader-specific reputation could be US President Trump's reputation for volatile and unpredictable foreign policy making. At the same time, the United States as a state maintains a strict reputation for nonnegotiation with nonstate groups who take US citizens as hostages. Although leader-specific reputations provide promising avenues for future research, here we focus on the foundational relationship between reputation for restraint and assurance credibility.

In addition, reputations are dynamic. Reputations decay as past behavior becomes less informative over time (Crescenzi 2018, 50), though the process of decay may be uneven, such that reputations persist until some shock leads states to adjust their beliefs (Dafoe, Renshon, and Huth 2014).[11] Conversely, reputations endure and strengthen if the relevant behavior is repeated, though a single salient action may be sufficient to ground a reputation.[12]

Turning from form to function, a reputation for restraint bolsters the credibility of a challenger's assurances for several reasons. First, challengers can use the cost of losing their reputation for restraint as evidence of their fidelity, "leveraging future payoffs as collateral to incentivize a commitment in the present" (Dafoe et al. 2014, 378). This mechanism operates when reputations hold instrumental value. For instance, Sartori (2005) argues that states treat diplomatic statements as informative because they prize a reputation for honesty, which reinforces the credibility of deterrent threats; the risk of future deterrence failure and war incentivizes states to pursue frank diplomacy in the present.[13] While Sartori is concerned with signals of resolve, analogous reasoning applies to assurances: as challengers may need to coerce adversaries in the future, they are incentivized to refrain from exploiting their opponent's concessions in the present.[14]

Second, a reputation for restraint signals that a state's true intentions are more likely to be benign. Weisiger and Yarhi-Milo (2015, 477) observe that a state's past behavior reveals its preferences over the issue at stake and over war in general, allowing opponents to discern whether a state has aggressive intentions or limited ambitions. A reputation for restraint suggests the latter, indicating that the challenger has historically not sought to press its advantage following a target's compliance.[15] By alleviating uncertainty over the challenger's intentions, a reputation for restraint mitigates the target's fear of future exploitation, increasing the credibility of the challenger's assurances.[16]

Though not tied to any specific case of coercion, an example of these dynamics can be found in the foreign policy making of German chancellor Otto von Bismarck. Following decisive victories against Austria and France, Bismarck refrained from annexing Austrian territory while skillfully committing Germany to a series of treaties and agreements, intentionally developing Germany's reputation as a restrained peacemaker in order to reassure regional powers that Germany sought no further revisions to the status quo. Though ultimately undermined by his successors, Bismarck's diplomacy temporarily dampened regional concerns about the German juggernaut, forestalling a counterbalancing alliance between Russia and France for several decades (Ullrich and Beech 2015).

Our argument on a reputation for restraint relates to Crescenzi's (2007, 2018) work on a "reputation for violence." Because past conflicts "can be considered evidence of failures to navigate crisis waters peacefully," Crescenzi argues that states with violent pasts can develop a reputation for hostility

11. That reputations may fade means that a reputation for restraint may be unable to remedy the fundamental, long-term insecurities that animate realist IR theory. Still, we argue that reputation does inform states' predictions about their adversaries' likely behavior in the near-term and shapes their cost-benefit analysis accordingly. For more on the temporal dynamics of reputation, see Harvey and Mitton (2016).

12. In this respect, our experimental design below offers a hard test for a theory of reputation building, as the reputation treatment involves only one episode of past behavior.

13. See also Trager (2016, 212–13; 2017, 20–21) for a discussion of "bargaining reputation."

14. Note that reputations for honesty and restraint are not synonymous. A state may be caught bluffing in a dispute (establishing a reputation for dishonesty), while nevertheless strictly abiding by all formal agreements it concludes (maintaining a reputation for restraint). States may also value a reputation for restraint for other reasons. For instance, a reputation for restraint may signal that a state is a responsible member of the "international community," a status that could convey benefits in various aspects of international politics, including alliances, participation in intergovernmental organizations and trade, and the avoidance of unwanted sanctions or diplomatic pressure.

15. Whether a reputation for restraint interacts with commitment problems as discussed by Powell (2006) and others depends on which reputation mechanism is engaged. Resolving uncertainty about a challenger's traits can avert conflicts arising from information problems, but not those arising from commitment problems, which may produce war even in the absence of uncertainty. If a reputation for restraint holds sufficient instrumental value for the challenger, however, it might nevertheless mitigate commitment problems by significantly depressing the challenger's payoff for reneging, rendering agreements enforceable.

16. Like any reputation, a reputation for restraint can emerge from a variety of inferential and behavioral microfoundations. Observers might conduct a naive extrapolation from past actions, or they might make more sophisticated inferences about unobserved factors like capabilities, worldview, or the cohesiveness of the ruling political coalition, which past behavior contributes to revealing. The specific microfoundations of reputation in international politics are beyond the scope of this article and therefore we reserve the issue for future research.

or incompetence during crises (Crescenzi 2007, 388–89; 2018, 78–79). Crescenzi shows that "violence begets violence," as states that acquire a reputation for violence are more likely to be involved in future militarized interstate disputes (MIDs). Setting aside the distinction between reputations for incompetence and restraint, we differ from Crescenzi in that we seek to explain the success or failure of coercion, not the onset of militarized disputes, which challengers may initiate irrespective of target behavior. While Crescenzi does not endeavor to develop the logic of assurances in coercion,[17] we provide a framework explicitly linking past behavior to coercive outcomes: a reputation for restraint encourages targets to acquiesce by mitigating their fears of future exploitation, both by signaling the challenger's limited ambitions and by incentivizing the challenger to uphold its commitment.

Importantly, we theorize that a reputation for restraint can be built without undermining the credibility of threats; one need not come at the cost of the other. It follows that, whereas relative power will have an indeterminate effect on coercive success, a reputation for restraint will have a positive net effect on the odds of coercive success.[18] A further implication of our argument is that reputational concerns matter for crisis bargaining even if a "reputation for resolve" is secondary to power as a determinant of threat credibility. Indeed, because power decreases assurance credibility, even the most powerful states cannot escape the value of establishing a reputation for restraint. Quite the opposite; their very strength makes a reputation for restraint even more important for coercive success.

This discussion produces the following hypotheses:

H3. All else equal, assurance credibility is increasing with the challenger's reputation for restraint.

H4. All else equal, a target's resolve to resist coercive demands decreases when the challenger maintains a reputation for restraint.

Table 1. Hypotheses

Variables	Threat Credibility	Assurance Credibility	Target Resolve
Power	+	−	Indeterminate
Rep. restraint	No effect	+	−

Note. Each cell represents the hypothesized direction of the direct effect of increasing each independent variable (left column) on each dependent variable (top row).

To conclude, relative power and a reputation for restraint jointly contribute to shaping the credibility of both the challenger's threats of punishment if met with defiance and its assurances against future demands if met with compliance. Taken together, their effects on the credibility of threats and assurances condition the target's resolve to resist the challenger's demands. Table 1 summarizes our expectations about how these factors condition threat credibility, assurance credibility, and the target's resolve.

EXPERIMENTAL DESIGN

In order to evaluate our claims about the causes and consequences of credible coercive assurances, we conduct a realistic, online, scenario-based survey experiment on a nationally representative sample of US citizens.[19] The survey—which was conducted in March 2018 and included 1,028 respondents—explored two causal relations: first, whether relative power and a reputation for restraint influence respondents' perceptions of the credibility of assurances as predicted; second, whether assurance credibility influences respondents' resolve to resist the challenger's demand in a realistic crisis scenario as predicted. A pre-analysis plan detailing the design and hypotheses was preregistered prior to fielding the survey.[20]

Survey methodology

Our research advances a growing body of international relations (IR) scholarship applying experimental methods to the study of international politics (Brutger and Kertzer 2018; Kertzer, Renshon, and Yarhi-Milo 2021; Tomz 2007; Tomz and Weeks 2013; Yarhi-Milo, Kertzer, and Renshon 2018). As discussed above, the primary benefit of this approach is that controlled crisis scenarios with randomized manipulations allow researchers to isolate the effects of key

17. It is sufficient for Crescenzi (2007) to assert that a reputation for violence makes states less trustworthy; as he writes, "a reputation for violence will increase the probability of the onset of new violence in a crisis, as states who perceive these reputations will have a harder time compromising in settlement attempts and trusting their opponents" (388–89).

18. Threat and assurance credibility covary only if driven by an inference relevant to both, such as a perceived lack of honor or integrity. In contrast, a primary contribution of this project is to show that some factors push threat and assurance credibility in opposing directions (e.g., relative power), while others manipulate one independent of the other (e.g., reputation for restraint). Thus it is theoretically possible for a state to have a poor reputation for following through on its threats while maintaining a good reputation for abiding by its assurances.

19. Survey sampling was conducted via Survey Sampling International (SSI). To test the viability of the hypothetical scenario, several pilot surveys were also conducted via Amazon's Mechanical Turk.

20. Evidence in Governance and Politics (EGAP) Study #20180108AA.

causal factors from the confounding variables or collinearity that undermine observational analyses. Moreover, exploring how the American public understands and reacts to crises offers valuable insights for IR scholars, given the numerous pathways through which public opinion has historically affected state behavior (Kertzer 2016, 50–51; Tomz and Weeks 2013; Tomz, Weeks, and Yarhi-Milo 2020). We therefore believe that a survey of the mass public permits reliable inference about state behavior.

At the same time, we recognize that the mass public and policy making elites are not identical populations. Due either to self-selection or experience, elites tend to be higher-level strategic thinkers than the average citizen. Specifically, elites are less prone to risk-acceptant behavior in the domain of losses, better able to reason into the future and consider their opponents' perspectives, and more patient (Hafner-Burton, Hughes, and Victor 2013; Hafner-Burton et al. 2014; LeVeck et al. 2014). While these findings do not imply that elites are hyper-rational game theorists immune to cognitive biases, they do lead Hafner-Burton et al. (2014, 866) to observe that scholars "may be on firmest ground when they assume that decision makers purposefully and carefully make choices with respect to what other decision makers might do, avoiding strategies that are strictly dominated." This is precisely the logic on which our theory depends—target states must first assess the risk of future exploitation and then decide whether to resist or accept the present demand.

One implication of these differences is that, compared to elites, the mass public may be less inclined to assess the problem of assurances (which requires strategic thought about the challenger's future behavior) and more likely to focus on the immediate threat. If so, our sample provides a harder test for our theory of assurance credibility and coercion relative to the prevailing wisdom, which privileges the credibility of threats. As a result, if our experiment reveals that assurance credibility affects resolve among the mass public as predicted by our theory, then it likely does so among elites as well. Here, we stress that we see no reason to believe that elites and the mass public would draw substantively distinct inferences about the relationships between our causal factors of interest (power and a reputation for restraint) and threat or assurance credibility. While it is possible that the magnitude of the effects could differ in an elite sample, their direction should not.

Taking stock, we accept that an ideal study would recruit an elite sample. Given the difficulty and cost of obtaining elite samples, however, our first priority is to develop the theory that will inform later research (Renshon, Lee, and Tingley 2017, 203). As Kertzer et al. (2021, 11–12) write, "Moving on to elite samples makes sense only after replications and extensions have increased our confidence in a given research program and narrowed down the plausible candidates for experimentation to a number of factors feasible for study in the extremely small samples that characterize elite experimentation." Still, to further alleviate concerns, we also check our analysis against a sample subset that more closely resembles the typical policy making elite, as suggested by Hafner-Burton et al. (2013) and Hafner-Burton et al. (2014) (see the online appendix). The results parallel those presented below.

Scenario design

Respondents read a vignette describing a hypothetical crisis between China and the United States occurring in 2025.[21] In the scenario, Beijing demands that the United States recognize Chinese sovereignty over the Diaoyu Islands, a historically contested territory that is currently administered by Japan (where they are referred to as the Senkaku Islands). The fictional Chinese leader assures the United States that he looks forward to peaceful relations but threatens to pursue "active measures" to protect China's territorial rights should the United States refuse this demand. Fictional US national security experts clarified this threat to mean that China would engage in provocations against US forces in the East China Sea, pressuring the United States to withdraw from the region. (The United States maintains a large military presence in the nearby island of Okinawa.)

These experts then offered additional, randomly manipulated information about our key independent variables. We manipulated relative power by varying whether the United States or China enjoyed military superiority in the region. To manipulate China's reputation for restraint, we told respondents that in 2020, China had concluded a similar territorial dispute with Vietnam over the Paracel Islands by agreeing to divide the island chain between them. Respondents then learned that during the intervening five-year period, China had either peacefully abided by that deal (priming a reputation for restraint) or that China had in the recent past pressured Vietnam to make additional concessions (priming the absence of a reputation for restraint). Following this description, respondents also viewed figure 1, a basic geographical representation of the scenario.

21. A hypothetical future scenario allows us to generate a realistic crisis while also providing some separation from respondents' prior beliefs about current US-China relations. Of course, these preconceptions undoubtedly influence to some degree how respondents interpret the scenario. Yet because our primes are randomly assigned, our analysis should not suffer from any systematic correlation between the treatments and unobserved confounding beliefs. As a result, these prior beliefs should not distort our inferences about the causal effects of power and reputation for restraint on respondents' resolve.

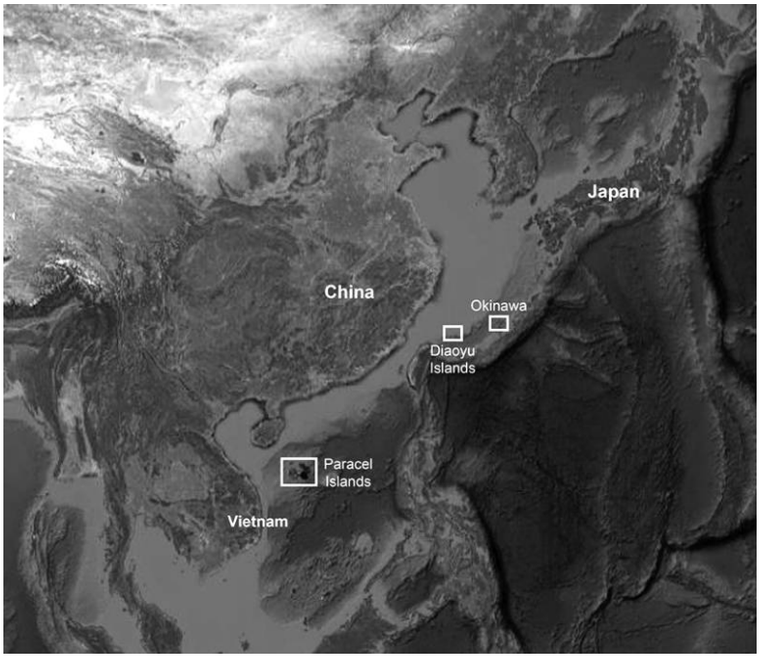

Figure 1. Scenario map

Respondents were asked to offer their views on three questions: the credibility of China's threats, that of its assurances, and their level of resolve vis-à-vis China's demands. Our main dependent variable was respondents' Resolve, which we measured by asking how they thought the United States should respond to China's demand for sovereignty over the Diaoyu Islands. Options ranged from "accept" to "stand firm regardless of the cost." As we understand resolve to be a function of the credibility of China's threats and assurances, we also included measures of these mediating variables. We measured Threat Credibility by asking respondents how probable it was that China would pressure US forces in the East China Sea if the United States refused China's demand. We measured Assurance Credibility by asking respondents how probable it was that China would pressure US forces even if Washington complied with its demand. Tables 2 and 3 summarize these main independent and dependent variables, as well as the control variables employed in the analysis below.

Our scenario asked respondents to assume the role of a policy advisor reacting to China's demand. This differs from the more common approach, which asks respondents to assume the role of a citizen approving or disapproving their leader's actions. We chose the advisor framework for several reasons. First, we wanted to focus respondents' attention on the factors implicated by our theory—resolve, threat credibility, and assurance credibility. As a citizen may approve or disapprove of a leader's actions for many other reasons, such as their perceptions of morality (Tomz and Weeks 2013) or the leader's partisanship (Trager and Vavreck 2011), we believe that the role of advisor is plausibly better suited for assessing theoretically relevant factors. In other words, to the extent that citizens' approval of their leader is affected by the

Table 2. Independent Variables

Variable	Level	Description
Balance of power	0: US superior	Experts say US has military superiority over China in region
	1: China superior	Experts say China has military superiority over US in region
Reputation	0: No reputation for restraint	China has pressed Vietnam to renegotiate past agreement
	1: Reputation for restraint	China has abided by past territorial agreement with Vietnam
Other controls		US president party, respondent party, education, income, political knowledge, gender, age, race, military experience, knowledge of foreign affairs

Table 3. Dependent Variables

Variable	Survey Question	Measure
Resolve	"How should the United States respond to China's demand for sovereignty over the Diaoyu Islands?"	4-point scale: *Accept* the demand; *Reject* the demand but ask Japan to concede if any conflict arises; *Reject* the demand but ask Japan to concede if more than 100 Americans die in conflict with China; *Reject* the demand at all costs
Assurance credibility	"Suppose that the United States concedes the Diaoyu Islands to China. Do you think that China will still attempt to contest U.S. military presence in the region at a later date?"	5-point likelihood scale
Threat credibility	"Suppose that the United States rejects China's demand for control of the Diaoyu Islands. Do you think that China will follow through on its threat to contest U.S. military presence in the region?"	5-point likelihood scale

issues we discuss, our advisor design captures the effect that is mediated by respondents' beliefs about optimal coercion. In addition, Kertzer and Renshon's (2015) preliminary investigation of hypothetical "perspective taking" speaks to external validity. The authors find that perspective taking can magnify respondents' preexisting inclinations about the use of force (making them "more like themselves") but find no evidence that it biased estimates of average treatment effects. Last, we think it valuable for the community of scholars to avoid an overly quick convergence on one method; diversity is important to preserve the gains from continued exploration of method. If the results from our study appear likely to be dependent on a particular set-up, future work can easily extend our investigation to other experimental designs.

RESULTS

We begin by establishing that power and a reputation for restraint affect assurance and threat credibility as predicted by our theory. Figure 2 presents the coefficients from several linear regressions of assurance and threat credibility on our two main causal factors, as well as their interaction (see tables 4 and 5). We find that increasing China's power significantly decreases the credibility of China's assurances against future aggression, while granting China a reputation for restraint significantly increases assurance credibility, supporting hypotheses 2 and 3. We also find that, in line with much prior scholarship, increasing China's power significantly increases the credibility of China's threat, affirming hypothesis 1. These findings are significant at least to the .95 level and grow stronger with the inclusion of controls.

Notably, a reputation for restraint does not appear to affect threat credibility, even as it strongly increases assurance credibility. This finding is also in line with our theoretical expectations (we theorized that reputation for restraint and threat credibility can be independent), and reinforces the intuition that a reputation for restraint enables coercers to reassure opponents without diminishing the credibility of threats—a particularly important feature for strong challengers. Together, these results support our claim that power and a reputation for restraint are important determinants of the credibility of threats and assurances and, through them, of the balance of coercion.

Several other features of these results are noteworthy. First, power is a stronger determinant of threat credibility than assurance credibility. Loosely, the power coefficients for assurance credibility are approximately half the size of the power coefficients for threat credibility. This difference in magnitude could obtain due to circumstances to our crisis scenario and could therefore vary across crises. In our case, respondents may have perceived power to be a more reliable indicator of threat credibility than assurance credibility because they focused on the reputation prime as another, more informative signal of assurance credibility. Alternative explanations for this difference currently lie beyond the scope of our theory, which is agnostic as to the relative magnitudes of the effects of power on threats and assurances. Second, we find no evidence of interaction effects between power and a reputation for restraint on either threat or assurance credibility, suggesting that these two causal factors operate independently.

We now turn to the consequences of assurance credibility by assessing whether power and a reputation for restraint ultimately affected coercive outcomes. Figure 3 presents the coefficients from the same regression specifications employed above, with Resolve as the dependent variable (see table 6). We first observe that granting China a reputation for restraint

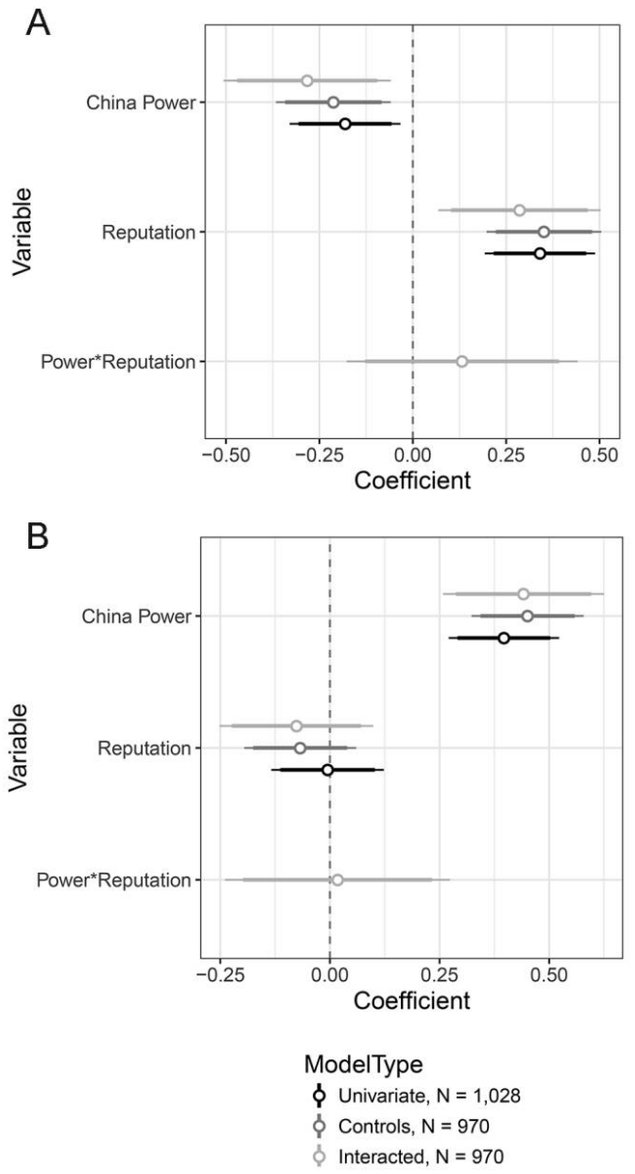

Figure 2. *A*, Assurance credibility. Respondents were asked how likely it was that China would contest US military presence in the East China Sea even if the United States agreed to its demands. Positive coefficients indicate greater confidence in China's assurances against such behavior, so we hypothesized a negative effect for China Power and a positive effect for China Reputation. For regression models, see table 4. *B*, Threat credibility. Respondents were asked how likely it was that China would contest US military presence in the East China Sea if the United States rejected its demand. Positive coefficients indicate greater confidence that China's threat was credible, so we hypothesized a positive effect for China Power. For regression models, see table 5.

significantly decreased respondents' willingness to resist China's demand, in line with hypothesis 4. Simulating predicted probabilities holding all other variables in our full model at their means, we find that granting China a reputation for restraint made respondents about 13% less willing to resist.[22] This result, significant at the .99 level, validates our core assertion: credible assurances improve the probability of coercive success.

In addition, we find that increasing China's power decreased respondents' willingness to resist China's demand. In other words, relative power boosted the odds of successful coercion. This result, which is consistent with the common understanding of the relationship between power, threat credibility, and coercive success, derives from the fact that, as discussed above, power preponderance increased the credibility of China's threats more than it decreased that of its assurances. Here, we reiterate that this dynamic is not preordained. In our survey, the median respondent was ambivalent about the credibility of China's threats, reporting that China was "as likely as not" to follow through (2 on a 0–4 scale). As a result, China had room to enhance its odds of success by increasing its threat credibility. It is possible, however, that when a challenger's baseline threat credibility is higher, increasing its power might offer only marginal improvements to threat credibility while substantially decreasing assurance credibility, thereby boosting the target's resolve and dampening the odds of coercive success.

Finally, we conducted a mediation analysis to isolate more precisely the causal mechanisms postulated by our theory. Mediation analysis provides a toolkit for evaluating whether the effects of power and a reputation for restraint on resolve

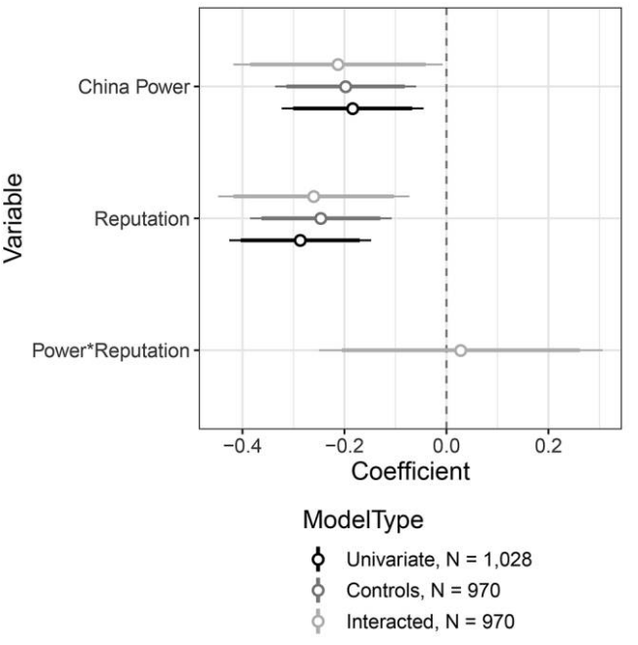

Figure 3. Resolve. Respondents were asked how the United States should respond to China's demand regarding the Diaoyu Islands. Positive values indicate increasing resolve to reject China's demand. We hypothesized a negative effect for China Reputation, while remaining agnostic on Power. For regression models, see table 6.

22. Simulations conducted using the Clarify/Zelig package.

Table 4. Assurance Credibility

	Dependent Variable: Assurance Credibility			
	If US Accepts Demand, Will China Still Contest US Military Presence?			
	(1)	(2)	(3)	(4)
China power	−.181*		−.213**	−.283*
	(.076)		(.078)	(.114)
China reputation		.340**	.351**	.286*
		(.075)	(.078)	(.111)
Power × reputation				.132
				(.158)
Controls			✓	✓
Constant	1.992**	1.720**	2.020**	2.059**
	(.053)	(.055)	(.174)	(.181)
Observations	1,028	1,028	970	970
R^2	.006	.019	.053	.053
Adjusted R^2	.005	.019	.038	.037
Residual SE	1.213 (df = 1026)	1.205 (df = 1026)	1.203 (df = 954)	1.203 (df = 953)
F-statistic	5.738* (df = 1; 1026)	20.405** (df = 1, 1,026)	3.525** (df = 15, 954)	3.348** (df = 16, 953)

Note. Robust SEs computed via Huber-White sandwich estimator. FE = fixed effects.
† $p < .1$.
* $p < .05$.
** $p < .01$.

Table 5. Threat Credibility

	Dependent Variable: Threat Credibility			
	If US Rejects Demand, Will China Contest US Military Presence?			
	(1)	(2)	(3)	(4)
China power	.397**		.451**	.442**
	(.064)		(.065)	(.094)
China reputation		−.005	−.068	−.076
		(.065)	(.065)	(.089)
Power × reputation				.017
				(.131)
Controls			✓	✓
Constant	1.784**	1.985**	2.263**	2.268**
	(.044)	(.047)	(.141)	(.147)
Observations	1,028	1,028	970	970
R^2	.036	.00001	.090	.090
Adjusted R^2	.035	−.001	.076	.075
Residual SE	1.030 (df = 1,026)	1.049 (df = 1,026)	1.009 (df = 954)	1.010 (df = 953)
F-statistic	38.120** (df = 1; 1026)	.007 (df = 1; 1026)	6.309** (df = 15; 954)	5.910** (df = 16; 953)

Note. Robust SEs computed via Huber-White sandwich estimator. FE = fixed effects.
† $p < .1$.
* $p < .05$.
** $p < .01$.

Table 6. Resolve

	Dependent Variable: Resolve			
	How Should US Respond to China's Demand?			
	(1)	(2)	(3)	(4)
China power	−.184**		−.198**	−.213*
	(.071)		(.071)	(.105)
China reputation		−.287**	−.246**	−.260**
		(.071)	(.071)	(.096)
Power × reputation				.028
				(.142)
Controls			✓	✓
Constant	1.823**	1.885**	.992**	1.000**
	(.049)	(.053)	(.152)	(.157)
Observations	1,028	1,028	970	970
R^2	.006	.016	.101	.101
Adjusted R^2	.006	.015	.087	.086
Residual SE	1.138 (df = 1026)	1.133 (df = 1026)	1.087 (df = 954)	1.088 (df = 953)
F statistic	6.711** (df = 1, 1,026)	16.384** (df = 1, 1,026)	7.138** (df = 15, 954)	6.687** (df = 16, 953)

Note. Robust SEs computed via Huber-White sandwich estimator. FE = fixed effects.
† $p < .1$.
* $p < .05$.
** $p < .01$.

flow through threat and assurance credibility as theorized. We proceed according to Imai and Yamamoto (2013), who propose a mediation estimator for multiple, causally dependent mediators.[23]

Figure 4 presents coefficient plots from mediation analyses conducted using the corresponding *mediation* package in R.[24] In the top row, we observe that although the total effect of power on resolve is negative, the indirect effect mediated by assurance credibility is positive and significant at the .95 level. This result is consistent with the argument that increasing the challenger's power has both positive and negative effects on the target's resolve—the former mediated through assurance credibility, the latter through threat credibility—though in this case the negative effect resulting from elevated threat credibility outweighed the positive effect from lessening assurance credibility. The bottom row demonstrates that the indirect effect of a reputation for restraint on resolve mediated through assurance credibility is negative and significant, consistent with the proposition that credible assurances are important for successful coercion.

DISCUSSION/CONCLUSION

The literature on international coercion, born in an era where the credibility of the threat of nuclear retaliation was under constant scrutiny, has long neglected the importance

23. On mediation techniques, see also Imai, Keele, and Tingley (2010) and Imai, Keele, and Yamamoto (2010). Mediation analysis relies on several assumptions, including that the mediators are conditionally independent of the outcome (there are no unobserved confounding variables) and that there is no interaction between the treatment and the mediators (the effect of a particular level of threat/assurance credibility on resolve does not vary based on the level of the power/reputation treatments). We see little reason to expect a mediator-treatment interaction, given that we find no power × reputation interaction effects (which would imply that certain combinations of the power and reputation primes encourage respondents to attend to threat and assurance credibility more carefully), although we cannot conclusively dismiss the possibility of unobserved confounding—on this latter problem, see also Dafoe, Zhang, and Caughey (2018). Future studies could attempt a factorial design as proposed by Acharya, Blackwell, and Sen (2018), though it will likely prove difficult to manipulate directly the threat and assurance credibility mediators without cueing respondents to the importance of these factors (violating what they call the manipulation exclusion restriction).

24. For mediation package documentation, see https://cran.r-project.org/web/packages/mediation/vignettes/mediation.pdf.

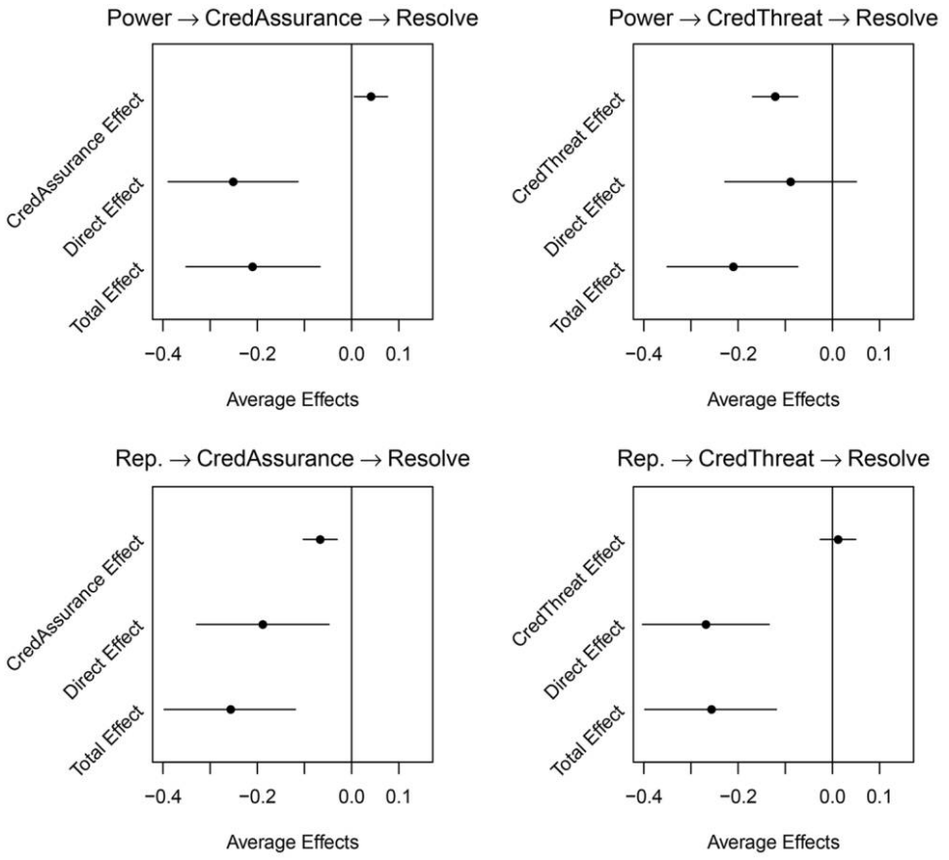

Figure 4. Each graph is a coefficient plot of a mediation regression analysis of resolve on power (*top row*) and reputation for restraint (*bottom row*), with the primary mediator of assurance credibility (*left column*) and threat credibility (*right column*), with 95% confidence intervals. For each *y*-axis, the uppermost coefficient is the average causal mediated effect of the stated variable on resolve, followed by the direct (i.e., nonmediated) effect, followed by the total effect.

of assurances for successful coercion. Armed with an incomplete understanding of coercion, scholars and policy makers alike have been ill equipped to explain some of the United States' most prominent coercive failures. In probing the factors that shape assurance credibility, this article joins a budding literature that we anticipate will grow into a rich field of study on assurances in crisis bargaining.

Our research suggests several promising avenues for further study. First, future scholarship should evaluate how the effects of power on threat and assurance credibility vary across different crisis contexts. The goal would be to identify the circumstances under which power increases threat credibility by more than it decreases assurance credibility, boosting the odds of successful coercion. While we found no evidence that power's deleterious effect on assurances was conditioned by the challenger's reputation for restraint, other factors may influence this relationship. Similarly, more research is necessary on the relationship between power and a reputation for restraint in conditioning assurance credibility. Is relative weakness sufficient to make assurances credible or do weak states, too, benefit from a reputation for restraint? Conversely, future work could ask the same question of the relationship between power and a reputation for resolve in shaping threat credibility.

Along these lines, future work should also consider the effects of other variables beyond power and a reputation for restraint on the credibility of assurances. To start, we should test whether the usual panoply of factors deemed to affect threat credibility—regime type, force posture, public statements, alliances, norms, and international institutions—also influence assurance credibility. Moreover, it is possible that "issue similarity," loosely understood as whether the issue at stake is similar to other issues the challenger might contest in the future, will influence assurance credibility. The intuition is that when issues that might be the object of future demands are similar to the current issue in dispute, targets may worry that present concessions will reveal their lack of resolve over future issues as well. As rational challengers would exploit this information by issuing subsequent demands, their assurances against future aggression become less credible. Thus, issue similarity may incentivize targets to resist now in order to deter future demands, triggering the reputational concerns that

are the source of coercive failure in Sechser (2010). In contrast, if present and possible future disputes are highly dissimilar, concessions would not provide much information about the target's future resolve, making backing down less risky for the target. Credible assurances are thus theorized to be especially important to coercion when issue similarity is high.

We conclude by reiterating that the study of assurances has a clear and pressing relevance for American foreign policy. Our theory suggests that the United States struggles with coercion not despite but rather because of its overwhelming strength, which, particularly when combined with an assertive foreign policy, makes it difficult to issue credible assurances against future demands. That dynamic is brought into stark relief by the Korean war example mentioned earlier, as US power projection led Mao to believe that US-China conflict was inevitable—perhaps Mao's thinking would have differed if the United States had sought to cultivate a clear reputation for restraint. More recently, some have argued that the incessant drive for regime change and the omnipresent threat of US aggression has hindered American nonproliferation efforts in both Iran and North Korea (Cebul 2015; Jervis and Rapp-Hooper 2018; Kydd 2018; Nowrouzzadeh, Pauly, and Rouhi 2017). Assurances seem especially relevant to US efforts to compel countries to refrain from acquiring nuclear weapons or, perhaps even more, to dispose of their nuclear arsenal. Yet assurances matter for coercive contexts beyond the nonproliferation regime. For example, the assurance problem may also be undermining US efforts to support democratizing movements in places like Venezuela, where the looming threat of human rights prosecution following democratization may dissuade autocrats and their military allies from relinquishing power.[25]

Broadly, then, the problem of how to boost the credibility of assurances against future punishment is a defining challenge for contemporary US coercive efforts. If the United States is to coerce its adversaries successfully in pursuit of its core policy interests, it must endeavor to credibly assure its opponents that its ambitions are limited and their vulnerability will not be exploited.

ACKNOWLEDGMENTS

We thank Michael Horowitz, Andrew Kydd, Jack Snyder, the editors and anonymous reviewers at the *Journal of Politics*, and the audience at the 2017 annual convention of the International Studies Association for helpful comments and guidance.

25. On the threat of post-tenure punishment and autocratic resistance to US pressure, see Escribà-Folch and Wright (2015). On Venezuela in particular, see Harris (2019) and Weiss (2019).

REFERENCES

Abrahms, Max. 2013. "The Credibility Paradox: Violence as a Double-Edged Sword in International Politics." *International Studies Quarterly* 57:660–71.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. "Analyzing Causal Mechanisms in Survey Experiments." *Political Analysis* 26:357–78.

Art, Robert J., and Patrick Cronin. 2003. *The United States and Coercive Diplomacy*. Washington, DC: US Institute of Peace Press.

Brutger, Ryan, and Joshua D. Kertzer. 2018. "A Dispositional Theory of Reputation Costs." *International Organization* 72:693–724.

Cebul, Matthew. 2015. "What the Iranian Nuclear Negotiations Tell Us about Soft Diplomatic Deadlines." Political Violence at a Glance. http://politicalviolenceataglance.org/2015/02/11/what-the-iranian-nuclear-negotiations-tells-us-about-soft-diplomatic-deadlines/.

Chamberlain, Dianne Pfundstein. 2016. *Cheap Threats: Why the United States Struggles to Coerce Weak States*. Washington, DC: Georgetown University Press.

Christensen, Thomas J. 1992. "Threats, Assurances, and the Last Chance for Peace: The Lessons of Mao's Korean War Telegrams." *International Security* 17 (1): 122–54.

Clare, Joe, and Vesna Danilovic. 2010. "Multiple Audiences and Reputation Building in International Conflicts." *Journal of Conflict Resolution* 54 (6): 860–82.

Crescenzi, Mark J. C. 2007. "Reputation and Interstate Conflict." *American Journal of Political Science* 51 (2): 382–96.

Crescenzi, Mark. 2018. *Of Friends and Foes: Reputation and Learning in International Politics*. Oxford: Oxford University Press.

Dafoe, Allan, and Devin Caughey. 2016. "Honor and War: Southern U.S. Presidents and the Effects of Concern for Reputation." *World Politics* 68 (2): 341–81.

Dafoe, Allan, Jonathan Renshon, and Paul Huth. 2014. "Reputation and Status as Motives for War." *Annual Review of Political Science* 14 (1): 371–93.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26:399–416.

Debs, Alexandre, and Nuno P. Monteiro. 2014. "Known Unknowns: Power Shifts, Uncertainty, and War." *International Organization* 68 (1): 1–31.

Debs, Alexandre, and Nuno P. Monteiro. 2016. *Nuclear Politics: The Strategic Logic of Proliferation*. Cambridge: Cambridge University Press.

Downes, Alexander B., and Todd S. Sechser. 2012. "The Illusion of Democratic Credibility." *International Organization* 66 (3): 457–89.

Escribà-Folch, Abel, and Joseph Wright. 2015. *Foreign Pressure and the Politics of Autocratic Survival*. Oxford: Oxford University Press.

Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88 (3): 577–92.

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

Fearon, James D. 1997. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *Journal of Conflict Resolution* 41 (1): 68–90.

Fuhrmann, Matthew, and Todd S. Sechser. 2014. "Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Deterrence." *American Journal of Political Science* 58 (4): 919–35.

Gelpi, Christopher, Peter D. Feaver, and Jason Reifler. 2006. "Success Matters: Casualty Sensitivity and the War in Iraq." *International Security* 30:7–46.

Gelpi, Christopher F., and Michael Griesdorf. 2001. "Winners or Losers? Democracies in International Crises, 1918–94." *American Political Science Review* 95 (3): 633–47.

Gibler, Douglas M. 2008. "The Costs of Reneging: Reputation and Alliance Formation." *Journal of Conflict Resolution* 52 (3): 426–54.

Guisinger, Alexandra, and Alastair Smith. 2002. "Honest Threats: The Interaction of Reputation and Political Institutions in International Crises." *Journal of Conflict Resolution* 46 (2): 175–200.

Hafner-Burton, Emilie M., D. Alex Hughes, and David G. Victor. 2013. "The Cognitive Revolution and the Political Psychology of Elite Decision Making." *Perspectives on Politics* 11:368–86.

Hafner-Burton, Emilie M., Brad L. LeVeck, David G. Victor, and James H. Fowler. 2014. "Decision Maker Preferences for International Legal Cooperation." *International Organization* 68:845–76.

Harris, Daniel. 2019. "After Bolton: A Dual Track Approach to Venezuelan Foreign Policy." Georgetown Security Studies Review. https://georgetownsecuritystudiesreview.org/2019/09/24/after-bolton-a-dual-track-approach-to-venezuelan-foreign-policy/#_edn27.

Harvey, Frank P., and John Mitton. 2016. *Fighting for Credibility: U.S. Reputation and International Politics*. Toronto: University of Toronto Press.

Hiroshima, Sean. 2015. "Divided Intentions: Iraqi Nuclear Weapons Policy between the First and Second Gulf Wars." Unpublished thesis, Stanford University. https://searchworks.stanford.edu/view/sc049qn8269.

Hopf, Ted. 1994. *Peripheral Visions: Deterrence Theory and American Foreign Policy in the Third World, 1965–1990*. Ann Arbor: University of Michigan Press.

Huth, Paul K. 1997. "Reputations and Deterrence: A Theoretical and Empirical Assessment." *Security Studies* 7 (1): 72–99.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15 (4): 309–34.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25 (1): 51–71.

Imai, Kosuke, and Teppei Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis* 21:141–71.

Jakobsen, Petter Viggo. 2012. "Reinterpreting Libya's WMD Turnaround: Bridging the Carrot-Coercion Divide." *Journal of Strategic Studies* 35:489–512.

Jentleson, Bruce W., and Christopher A. Whytock. 2005/6. "Who 'Won' Libya? The Force-Diplomacy Debate and Its Implications for Theory and Policy." *International Security* 30:47–86.

Jervis, Robert. 2003. "The Confrontation between Iraq and the U.S.: Implications for the Theory and Practice of Deterrence." *European Journal of International Relations* 9 (2): 315–37.

Jervis, Robert, and Mira Rapp-Hooper. 2018. "Perception and Misperception on the Korean Peninsula." *Foreign Affairs*. https://www.foreignaffairs.com/articles/north-korea/2018-04-05/perception-and-misperception-korean-peninsula.

Kertzer, Joshua D. 2016. *Resolve in International Politics*. Princeton, NJ: Princeton University Press.

Kertzer, Joshua D., and Jonathan Renshon. 2015. "Putting Things in Perspective: Mental Stimulation in Experimental Political Science." Working paper. http://people.fas.harvard.edu/jkertzer/Research_files/PT%20Web%20version.pdf.

Kertzer, Joshua D., Jonathan Renshon, and Keren Yarhi-Milo. 2021. "How Do Observers Assess Resolve?" *British Journal of Political Science* 51 (1): 308–30.

Knopf, Jeffrey W. 2012a. "Introduction." In Jeffrey W. Knopf, ed., *Security Assurances and Nuclear Nonproliferation*. Stanford, CA: Stanford University Press, 1–13.

Knopf, Jeffrey W. 2012b. "Varieties of Assurance." *Journal of Strategic Studies* 35 (3): 375–99.

Kydd, Andrew. 2000. "Trust, Reassurance, and Cooperation." *International Organization* 54 (2): 325–57.

Kydd, Andrew. 2018. "Promises on North Korea Are Easy to Make but Hard to Keep: Here's Why." The Monkey Cage. https://www.washingtonpost.com/news/monkey-cage/wp/2018/06/07/promises-on-north-korea-are-easy-to-make-but-hard-to-keep-heres-why/?utm_term=.d56f3dd49072.

Kydd, Andrew H., and Roseanne McManus. 2015. "Threats and Assurances in Crisis Bargaining." *Journal of Conflict Resolution* 61 (2): 1–24.

Lake, David A. 2010. "Two Cheers for Bargaining Theory: Assessing Rationalist Explanations of the Iraq War." *International Security* 35:7–52.

Lanoszka, Alexander. 2018. *Atomic Assurance: The Alliance Politics of Nuclear Proliferation*. Ithaca, NY: Cornell University Press.

LeVeck, Brad L., D. Alex Hughes, James H. Fowler, Emilie Hafner-Burton, and David G. Victor. 2014. "The Role of Self-Interest in Elite Bargaining." *PNAS* 111:18536–41.

Levy, Jack S., Michael K. McKoy, Paul Poast, and Geoffrey P. R. Wallace. 2015. "Backing Out or Backing In? Commitment and Consistency in Audience Costs Theory." *American Journal of Political Science* 59:988–1001.

Mercer, Jonathan. 1996. *Reputation and International Politics*. Ithaca, NY: Cornell University Press.

Midford, Paul. 2010. "The Logic of Reassurance and Japan's Grand Strategy." *Security Studies* 11 (3): 1–43.

Monteiro, Nuno P. 2009. "Three Essays on Unipolarity." PhD thesis, University of Chicago.

Monteiro, Nuno P. 2014. *Theory of Unipolar Politics*. Cambridge: Cambridge University Press.

Morrow, James D. 1989. "Capabilities, Uncertainty, and Resolve: A Limited Information Model of Crisis Bargaining." *American Journal of Political Science* 33 (4): 941–72.

Myerson, Roger B. 2006. "Force and Restraint in Strategic Deterrence: A Game-Theorist's Perspective." Unpublished paper based on a talk presented at the Chicago Humanities Festival on Peace and War. https://www.globalsecurity.org/military/library/report/2007/ssi_myerson.pdf.

Nowrouzzadeh, Sahar, Reid Pauly, and Masha Rouhi. 2017. "This Is Why Trump's Strategy for Iran Will Fail." *National Interest*. http://nationalinterest.org/feature/why-trumps-strategy-iran-will-fail-23748.

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

Pauly, Reid. 2019. "'Stop or I'll Shoot, Comply and I Won't': The Dilemma of Coercive Assurance in International Politics." Dissertation, MIT.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60 (1): 169–203.

Press, Daryl. 2004/5. "The Credibility of Power: Assessing Threats during the 'Appeasement' Crises of the 1930s." *International Security* 29 (3): 136–69.

Press, Daryl. 2005. *Calculating Credibility: How Leaders Assess Military Threats*. Ithaca, NY: Cornell University Press.

Renshon, Jonathan, Allan Dafoe, and Paul Huth. 2018. "Leader Influence and Reputation Formation in World Politics." *American Journal of Political Science* 62:325–39.

Renshon, Jonathan, Julia J. Lee, and Dustin Tingley. 2017. "Emotions and the Micro-foundations of Commitment Problems." *International Organization* 71:189–218.

Sartori, Anne E. 2005. *Deterrence by Diplomacy*. Princeton, NJ: Princeton University Press.

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

Sechser, Todd S. 2010. "Goliath's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64 (4): 627–60.

Sechser, Todd S. 2016. "Reputations and Signaling in Coercive Bargaining." *Journal of Conflict Resolution* 62 (2): 318–45.

Snyder, Glenn, and Paul Diesing. 1977. *Conflict among Nations: Bargaining, Decision Making, and System Structure in International Crises*. Princeton, NJ: Princeton University Press.

Snyder, Jack, and Erica Borghard. 2011. "The Cost of Empty Threats: A Penny, Not a Pound." *American Political Science Review* 105 (3): 437–56.

Stam, Allan C. 1996. *Win, Lose, or Draw: Domestic Politics and the Crucible of War*. Ann Arbor: University of Michigan Press.

Stein, Janice. 1991. "Reassurance in International Conflict Management." *Political Science Quarterly* 106 (3): 431–51.

Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61 (4): 821–40.

Tomz, Michael., and Jessica Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107:693–724.

Tomz, Michael, Jessica Weeks, and Keren Yarhi-Milo. 2020. "Public Opinion and Decisions about Military Force in Democracies." *International Organization* 74:119–43.

Trager, Robert F. 2016. "The Diplomacy of War and Peace." *Annual Review of Political Science* 19:205–28.

Trager, Robert F. 2017. *Diplomacy: Communication and the Origins of International Order*. Cambridge: Cambridge University Press.

Trager, Robert F., and Lynn Vavreck. 2011. "The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party." *American Journal of Political Science* 55:526–45.

Ullrich, Volker, and Timothy Beech. 2015. *Bismarck: The Iron Chancellor*. London: Haus.

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

Weisiger, Alex, and Keren Yarhi-Milo. 2015. "Revisiting Reputation: How Past Actions Matter in International Politics." *International Organization* 69 (2): 473–95.

Weiss, Laura. 2019. "U.S. Sanctions, and an Embargo, Will Only Worsen Venezuela's Humanitarian Crisis." *World Politics Review*. https://www.worldpoliticsreview.com/trend-lines/28097/u-s-sanctions-and-an-embargo-will-only-worsen-venezuela-s-humanitarian-crisis.

Wolford, Scott. 2019. *The Politics of the First World War: A Course in Game Theory and International Security*. Cambridge: Cambridge University Press.

Yarhi-Milo, Keren, Joshua D. Kertzer, and Jonathan Renshon. 2018. "Tying Hands, Sinking Costs, and Leader Attributes." *Journal of Conflict Resolution* 62:2150–79.