Constraining to Coerce*

Livio Di Lonardo† Scott A. Tyson‡

Abstract

Governments use a variety of tools to discourage, impede, or limit the ability of foreign adversaries to alter the status quo. Some of these measures seek to constrain an opponent's capacity to hurt, while others seek to coerce an opponent to take (or not) a particular action. We develop a theory to study how constraining and coercive measures interact strategically. Building on canonical models of deterrence, we first identify when coercive measures, in isolation, can curb transgressions from an aggressor. We then identify conditions when coercive and constraining measures are substitutes and when they are complements. In some cases, constraining measures can make deterrence via coercive measures possible, when they would fail completely in isolation. Our results also offer insights about measuring the effectiveness of various diplomatic tools, in particular economic sanctions. We highlight a series of novel empirical challenges stemming from the interaction of ecological effects and selection effects.

*We thank Ashley Anderson, Anna Bassi, Tyson Chatagnier, Richard DiSalvo, Jessica Gottlieb, Seth Hill, Federica Izzo, David Lake, Massimo Morelli, Santiago Olivella, Pablo Pinto, Peter Schram, Mehdi Shadmehr, Brad Smith, Gergely Ujhelyi, participants at the the 2022 Alghero Political Economy Conference, the Integrated Speaker Series and the University of California in San Diego, the Formal and Quantitative Seminar at the University of North Carolina, and University of Houston Hobby School Speaker Series for invaluable comments and discussions.

†Assistant Professor, Department of Social and Political Sciences and Carlo F. Dondena Centre for Research on Social Dynamics and Public Policy, Bocconi University. Contact information: livio.dilonardo@unibocconi.it

‡Associate Professor, Department of Political Science, University of Rochester, and Research Associate, W. Allen Wallis Institute of Political Economy, University of Rochester. styson2@ur.rochester.edu.

In the last months of 2021, amid an alarming buildup near their shared border, Ukrainian officials faced the prospect of an imminent Russian invasion. At that time, the main task at hand for them—along with their allies—was to avoid or mitigate Russian aggression. President Biden ruled out the possibility of putting U.S. combat troops in Ukraine, but repeatedly pledged to impose unprecedented economic sanctions if Russia invaded. The idea was that the threat of sanctions, as an alternative to military force, could deter Russia. Ukrainian President, Volodymyr Zelenskiy, instead urged Western countries to implement economic sanctions before Russia invades, arguing that doing so would prevent aggression, or at least limit its scope, by constraining Russia's capabilities. US Secretary of State, Antony Blinken, in an interview with Dana Bash,[1] argued instead that "when it comes to sanctions, the purpose of those sanctions is to deter Russian aggression. So if they're triggered (before the aggression), you lose the deterrent effect."

Zelenskiy and Blinken agreed about the underlying motives and resolve of Vladimir Putin and the Russian military. They disagreed about whether economic sanctions should be used as a constraining measure or as a coercive threat. Blinken's argument was that, in light of the US stated position of "no boots on the ground," no other threatened military action could help Ukraine and its allies achieve deterrence. The only hope to deter Russian aggression was the threat of economic sanctions, and employing them preemptively would destroy that hope. Zelenskiy's argument was that constraining Russian capabilities was the only hope at reducing transgressions.

In this paper we argue that constraining measures can improve the ability to deter transgressions in a scenario where coercive measures, by themselves, would not be able to do so. Rather than stressing the usual distinction between military actions and economics sanctions, predicated exclusively on the presence (or lack of) force, we focus on the distinction

[1]Transcript: https://www.state.gov/secretary-antony-j-blinken-with-dana-bash-of-cnn-state-of-the-union-2/

between constraining and coercive measures more generally. This distinction reflects recent insights that some military actions—like missile strikes, no-fly zones, or low-level military operations like "hassling"— constrain more than coerce an opponents ability to project force (Schram 2021, 2022; Joseph 2023), and how economic sanctions have been used both to constrain and coerce adversaries (Spaniel and Smith 2015; McCormack and Pascoe 2017; Kavaklı, Chatagnier and Hatipoğlu 2020; Grillo and Nicolò 2023). This distinction centers our analysis around the *strategic* incentives that different foreign policy tools engender among target countries.

We develop a theory that focuses on the relationship between a defender and an aggressor, along the lines of common theoretical formulations that are typically used to study deterrence (e.g., Powell 1987, 1989, 1990; Wagner 1992; Chassang and Padró i Miquel 2010; Kydd and McManus 2017; Baliga, Bueno de Mesquita and Wolitzky 2020; Di Lonardo and Tyson 2022). In our model, *Defender* can impose constraining measures, which include things like financial sanctions and tactical strikes. Following the implementation (or not) of constraining measures, *Aggressor* chooses whether to challenge a status quo, and if so, by how much, meaning she decides how severe of a transgression to pursue. Following a transgression, of whatever size, Defender has the chance to retaliate, which although costly to Defender, imposes punitive costs on Aggressor. Retaliation can take the form of economic sanctions, such as trade embargoes or military actions.

Our model departs from standard formulations of deterrence in two ways. First, by introducing constraining measures, and second, to measure the impact of constraining measures, allowing Aggressor to choose the severity of her transgression (rather than a binary choice, such as escalate or not). We first show that the threat of retaliation generates a reduction in transgressions, which we call *the deterrence effect*, under two well-known conditions. First is the *credibility* of the threat, which corresponds to identifying when Defender finds retaliation incentive compatible at the point she needs to carry it out. The second is the *capability* of

the threat, and refers to ensuring that a retaliatory action, if carried out, is severe enough that Aggressor reduces transgressions to avoid retaliation. We show that the deterrence effect only arises for a certain range of transgressions, which we call the *deterrence range*. The deterrence effect and the deterrence range highlight an important conceptual distinction between the *intensive margin* and *extensive margin* of deterrence respectively. The deterrence range (extensive margin) isolates when the deterrence effect is present, i.e., nonzero, while the deterrence effect (intensive margin) identifies how much of the reduction in transgressions (if present) are attributable to the retaliatory threat. We show that deterrence's intensive and extensive margins depend on the same factors, and sometimes in ways that are not reinforcing. This shows that the empirical evaluation of deterrence is more involved than is typically acknowledged, since it involves concerns distinct from selection bias, but related to ecological issues about a sample of country dyads. Our results suggest a reorientation of the empirical assessment of deterrence, focusing instead on carefully articulating both of its intensive and extensive margins, and considering how they each will effect one's sample as well as outcomes (on average).

Our main analysis centers on how constraining and coercive policies work together and highlights how, depending on the initial capacity of an aggressor to transgress, defender uses constraining and coercive actions differently. Specifically, we divide our analysis between "weak" and "strong" aggressors. When Aggressor is weak, deterrence and constraining measures are never used together. Specifically, we show that in any equilibrium either the deterrence effect, or a similarly defined constraining effect, are positive—but never both. In particular, there are two possibilities. First, when constraining measures lack effectiveness—or are overly costly—then no constraining measures are used, and the (strictly positive) deterrence effect is uniquely responsible for any reduction in transgressions. Second, when constraining measures are highly effective (relative to their cost), then they are used so heavily that Aggressor becomes constrained to the point that retaliation loses credibility for

Defender. Thus, our model provides a theoretical foundation for the common intuition that countries choose *between* different diplomatic tools, and shows that this intuition is implicitly predicated on a power imbalance. Critically, we do not *assume* that constraining measures and deterrence are substitutes, but instead, establish that they do not appear together when analysis is restricted to weak aggressors (i.e., particular country pairs).

When aggressor is strong, the consequences of retaliation are not severe enough to activate deterrence. In this case, some amount of constraining measures are used in any equilibrium, and how much depends on their effectiveness relative to their cost. If constraining measures are highly effective, or highly ineffective (relative to their cost), then they are the only tool employed by Defender. In both scenarios, the threat of retaliation cannot deter (for different reasons), and the full reduction in transgressions is achieved solely through constraining measures.

The most novel and interesting case is when an aggressor is strong and constraining measures are not too costly at lower levels, but where their extended use is overly arduous. In this case, constraining measures and deterrent threats are used together—as *complements*. In particular, Defender uses constraining measures to reduce Aggressor's initial capability to transgress just enough that deterrence *becomes* effective. This result is important because it identifies how constraining measures serve to promote the effectiveness of deterrence, thus causing an enhanced and discontinuous reduction in transgressions. Thus countries can *activate* deterrence by implementing constraining measures against a strong aggressor, implying that constraining and coercive actions are used in conjunction to curb aggression from aggressors who would otherwise be more hostile members of the international community.

We conclude by discussing the empirical implications of our theory about the strategic relationship between constraining and coercive actions, focusing in particular on the insights our theory can offer for empirically assessing the effectiveness of such tools. We show that, absent special circumstances, such as a natural experiment, the effectiveness of constraining

measures cannot be assessed because deterrence introduces bias. Second, we document distinct sources of bias in empirical comparisons between cases where constraining measures are used and those in which they are not. We offer possible solutions that range from sample restrictions to methodological techniques.

Related Literature

Our framework seeks to clarify the strategic connection between the two types of actions governments often use to induce compliance from other countries, namely constraining and coercive measures. The theoretical literature that features both constraining and coercion, either in isolation or together, has focused on three key mechanisms: coercing, signaling, and constraining.[2]

For decades, scholars have enquired when and how the threat of military actions and/or economic sanctions can coerce opponents (e.g., Schelling 1960, 1966). Along with establishing the centrality of the credibility and capabilities of such threats, recent work has analyzed how a defender's ability to deter is affected by problems of attribution (Baliga, Bueno de Mesquita and Wolitzky 2020), entering secret alliances (Bils and Smith 2023), confronting an opponent who might be unappeasable (Gurantz and Hirsch 2017), one who faces internal political challenges (Di Lonardo and Tyson 2022), and strategic ambiguity over military capabilities (Baliga and Sjöström 2008).

Closer to our setup, Schram (2021, 2022), and Joseph (2023) consider the possibility that defenders can preemptively engage in low-level military operations, or "hassling," which are short of war, to determine whether such actions can deter *by themselves*.[3] Joseph (2023)

[2]For broader summaries of the theoretical literature in international politics see Jackson and Morelli (2011) and Ramsay (2017). For theories of diplomacy using mechanism design to understand mediation, see Hörner, Morelli and Squintani (2015) and Meirowitz, Morelli, Ramsay and Squintani (2019).

[3]Spaniel and İdrisoğlu (2023) analyze how the availability of different war strategies, which vary in effectiveness and costliness, affect bargaining outcomes. They show that as the more effective option becomes more costly, war becomes more likely due to a substitution effect away from the more effective option towards

shows how low-level military actions can help deter challengers from developing new coercive technologies. Along with several differences in modeling assumptions, the fundamental difference between their results and ours is that they focus on challengers who invest in higher capabilities and defender's hassling to degrade challenger's capabilities. Consequently, a defender cannot preemptively constrain, but use the threat of hassling to deter investments (i.e., deterrence by denial).

Alternatively, certain military actions and/or economic sanctions, taken preemptively, can function as a signal of resolve, i.e., a commitment to not back down in the event of an escalated conflict. Uncertainty over a state's resolve has long been identified as one of the main drivers of conflicts (Fearon 1995), which is why early in their tenure, when this uncertainty is particularly severe, leaders are more likely to engage in a conflict (Wolford 2007). For similar reasons, conflicts may last longer (Smith and Spaniel 2019), or more economic interdependence may not necessarily be a force for peace (Spaniel and Malone 2019). Precisely because uncertainty can be a powerful source of conflict, leaders might issue military threats in an attempt to signal high resolve to their opponents during bargaining, in order to achieve better deals (Slantchev 2011). However, signaling resolve via costly actions might turn out to be more inefficient than war when signaling is not possible (Spaniel 2021), and it might come at the expense of support from coalition partners concerned about the cost of a conflict (Wolford 2014).

The logic of constraining has mostly been applied to economic sanctions, both in terms of the ability to project force immediately or in terms of developing new technologies. Many studies have stressed the constraining function of sanctions (e.g., Clark and Reed 2005; Giumelli 2011), focusing on how they can reduce the severity of the transgression an aggressor eventually chooses. McCormack and Pascoe (2017) develop a dynamic bargaining game where economic sanctions can alleviate potential power shifts that lead to preven-

the less effective and less costly one.

tive wars. They also show how very effective sanctions can trigger a war, by generating a commitment problem for the target state. The effectiveness of sanctions at degrading capacity has been shown to hinge on several factors, including whether the sender has a comparative advantage in goods exported to the target (Kavaklı, Chatagnier and Hatipoğlu 2020); on the target's elasticities of demand and substitution (Kustra 2023); the share of the target's market retained by sender's firms relative to its foreign competitors (Bapat and Kwon 2015); and whether the sender has the support of the target's major trading partners (McLean and Whang 2010). Differentiating among different types of sanctions, several scholars have argued that smart or targeted sanctions are more effective than comprehensive sanctions at achieving the political goals that led to their application. Baliga and Sjöström (2022) show instead how increasing targeted sanctions can reduce the effectiveness of comprehensive sanctions. Sanctions have also been shown to have a destabilizing effect on some regimes (Marinov 2005; Grauvogel, Licht and von Soest 2017), to affect the degree of power-sharing between authoritarian leaders and the elites supporting them (Grauvogel, Marinov and Wong 2022), and their imposition is more frequent when senders face uncertainty about the targeted leader's grip on power (Spaniel and Smith 2015).

Our framework contributes by advancing our understanding of when actions that constrain an adversary, and actions that coerce an adversary, are substitutes or complements. While some empirical studies have noted that such tools are often employed together (Pape 1997; Giumelli 2011), or one after the other (Lektzian and Sprecher 2007), but the strategic reasons why have not been articulated. We formulate such reasons in a strategic framework, and place particular emphasis on the insights our model gives for how to improve the empirical assessment of the effectiveness of different diplomatic tools.

Model

We consider the strategic interaction between two countries: Defender, D, and Aggressor, A. In the first stage of the game, Defender chooses a level of constraining measures, $x \in \mathbb{R}_+$, to impose on Aggressor. Imposing constraints is costly for Defender, and the cost is captured by the function $c(x)$, which is smooth, strictly convex, and strictly increasing, i.e., $c'(x) > 0$, and is normalized so that $c(0) = 0$ and $\lim_{x\to+\infty} c'(x) = +\infty$.

After observing the level of constraining measures, x, Aggressor chooses a transgression level $\pi \in [0, \overline{\pi}(x)]$. The size of the transgression cannot exceed $\overline{\pi}(x)$, which is a smooth, strictly convex, and strictly decreasing function of the level of constraining measures, x. The maximum transgression, $\overline{\pi}(x)$, depends on a number of substantive features that are outside of our model, such as physical constraints or domestic aspects of A (political or otherwise), that make potential transgressions exceeding $\overline{\pi}(x)$ either impossible or undesirable.[4] For example, economic sanctions that prevent trade with A can make more severe transgressions no longer feasible by making some weapon materials unavailable, prohibitively costly, or too risky/dangerous to develop. Similarly, tactical strikes against A that destroy support equipment or infrastructure restrict A's ability to project force.

Defender observes the size of the transgression, π, and chooses whether to retaliate, a choice denoted by $r \in \{0, 1\}$, where $r = 1$ corresponds to retaliation and $r = 0$ is not retaliate. Retaliation is costly both for Defender to carry out, captured by $k_D > 0$, and for Aggressor to endure, captured by $k_A > 0$. Retaliation also reduces the severity of transgressions, captured by $q \in [0, 1)$, which reduces the severity of a transgression of size π to $(1 - q)\pi$.[5]

A transgression is beneficial for Aggressor and costly for Defender. The expected payoff

[4] The latter is easily microfounded via a more general framework where transgressing is costly and $\overline{\pi}(x)$ is obtained as the size of the transgression that maximizes Aggressor's net benefit.

[5] This assumption is standard and is needed in our setting to ensure that retaliation is incentive compatible for Defender.

for Defender is

$$U_D(\pi, r, q) = -(1 - qr)\pi - r \cdot k_D - c(x), \tag{1}$$

and for Aggressor is

$$U_A(\pi, r, q) = (1 - qr)\pi - r \cdot k_A. \tag{2}$$

Retaliation is costly to both Defender and Aggressor, and it is this that serves to partially align D and A's interests in our model.

To summarize, the timing of the game is as follows: 1) Defender chooses a level of constraining measures, x; 2) Aggressor observes x and chooses a level of transgression, π; 3) Defender observes π and chooses whether to retaliate, r. Our solution concept is pure strategy subgame perfect Nash equilibrium.

It is natural to interpret $\overline{\pi}(0)$, the maximum level of transgressions A can pursue when $x = 0$, as measuring the initial balance of strength (or power) between Aggressor and Defender. It determines the initial capacity of A to transgress against D's interests, and so higher $\overline{\pi}(0)$ (pointwise) corresponds to a more powerful Aggressor (initially) vis-a-vis Defender. Constraining measures restrict the transgressions Aggressor can pursue, which highlights how such actions alter the balance of power between an aggressor and defender, and thus, have the ability to influence downstream crises.

Comments on the Model

Before proceeding to the main analysis, we discuss the interpretation of some key features of our model. Our theory focuses on the relationship between constraining measures that preempt an aggressor's opportunity to transgress as well as actions that are pursued in response to transgressions. Our model is intentionally constructed to maintain a tight connection with standard formulations of deterrence and international crises (e.g., Powell 1987, 1989, 1990; Fearon 1994; Schultz 1998; Di Lonardo and Tyson 2022). To maintain conceptual clarity, we

intentionally omit many important factors of international politics that, although important, are not key to understanding the relationship between constraining measures and retaliatory actions (the focus of our study). Importantly, omitting such factors is equivalent to holding them fixed in an empirical analysis (Paine and Tyson 2020).

The actual mechanisms that constrain an aggressor, such as missile strikes that debilitate an army's ability to advance (e.g., taking out anti-aircraft weapons), or increases in the power of domestic threats that alter the domestic political calculus of the regime, thus leading $\overline{\pi}(x)$ to be decreasing, are beyond the scope of our model, which focuses on their strategic influence. Targeted assassinations are well-known as a policy used by The Israeli Institute for Intelligence and Special Operations, otherwise known as Mossad (Bergman 2018). For example, Operation Damocles was a campaign targeting German scientists working to build rockets for the Egyptian military in the early 1960s. The purpose of the operation was to constrain the Egyptian military's capabilities in projecting force.[6]

In our model, military actions and economic sanctions can manifest at two different times: before and after the transgression choice. This serves to distinguish actions that preemptively constrain an aggressor's capacity from those that change aggressor's cost-benefit calculus. Another important dimension is the time it takes for such actions to mature. For instance, the full effect of some kinds of economic sanctions take time to materialize, whereas others, like monetary sanctions, have an effect essentially overnight, potentially crippling a target's currency (Kirshner 1997*a*). Similarly, some military strikes are implemented in a manner of hours (e.g., drone strikes) whereas a full-scale intervention takes months. Although the time it takes any action, whether constraining, punitive, or transgressive, to mature is important, it is beyond the scope of our model, but may be helpful for distinguishing the strategic role of any concrete tool.

There are many examples where economic sanctions were used punitively. For instance,

[6]It was viewed as successful until Egypt ultimately purchased Soviet Scud missiles.

in 1987, led by the United Nations, several countries placed an oil embargo, among other trade sanctions, on South Africa. The goal of these sanctions were to end the institutional system of racial segregation known as apartheid, and gain the release of Nelson Mandela from prison.[7] By the late 1980s, the collection of sanctions against South Africa had led to extreme inflation and capital flight—a cost political leaders could not ignore. Beginning in 1990, and following the release of Nelson Mandela on February 11, negotiations began the process of dismantling aparthaid in South Africa.

Finally, in our model there is a single Defender, who moves twice, first by choosing constraining measures and then later when deciding whether to retaliate in response to transgressions. This structure, although capturing many substantive cases, does not directly reflect cases where a different actor employs constraining measures and another retaliates. In an extension, developed in Supplemental Appendix A, we consider this possibility and show that the key strategic forces we highlight remain, where their magnitude depends on the extent to which different Defenders dislike transgressions differently.

The Deterrence Effect

We begin our analysis by focusing on the influence of retaliation threats by Defender. Retaliation could take different forms, from military actions (e.g., boots on the ground) to economic sanctions (e.g., trade embargoes). The analysis of this section focuses on deterrence, starting with the subgame that begins at Aggressor's transgression decision (the second stage). At this point, the level of constraining measures chosen by Defender, x, is exogenously fixed, and hence, $\overline{\pi} \equiv \overline{\pi}(x)$.

Proceeding backward, in the last stage of the game, if Aggressor has transgressed to level

[7] In the United States, the Anti-Apartheid Act was vetoed by President Reagan, and overriden by Congress, an effort led by Indiana Senator Richard Lugar (R).

π, then retaliation is sequentially rational for Defender whenever

$$-k_D - (1-q)\pi \geq -\pi$$

which rearranges to

$$k_D \leq q\pi. \tag{3}$$

This is the *credibility constraint*, and it represents when retaliation by D is a credible (i.e., sequentially rational) response to a transgression of size π by A.

Lemma 1 *Defender's retaliation decision rule is*

$$r^*(\pi) = \begin{cases} 1 & \text{if } \pi > \frac{k_D}{q} \\ 0 & \text{otherwise .} \end{cases}$$

Defender's decision rule is to retaliate if A has chosen a "severe" enough transgression.

Given D's response to different levels of transgressions, we move to Aggressor's decision problem. By sequential rationality, and since the maximum transgression is $\overline{\pi}$, we can write Aggressor's problem as

$$\max_{\pi \in [0,\overline{\pi}]} \pi(1 - r^*(\pi)) + r^*(\pi)((1-q)\pi - k_A).$$

Knowing how Defender will react to every transgression, Aggressor can choose a transgression that will not trigger retaliation, or choose a higher transgression that will trigger Defender's retaliation. The transgression that makes Aggressor indifferent between enduring retaliation and not, denoted by $\hat{\pi}$, is that which satisfies

$$\frac{k_D}{q} = (1-q)\hat{\pi} - k_A,$$

which, after rearranging, becomes,

$$\hat{\pi} = \frac{k_D + qk_A}{q(1-q)}. \tag{4}$$

For a transgression slightly lower than $\hat{\pi}$, because it would trigger retaliation, Aggressor is better off at $\pi = \frac{k_D}{q}$. For a transgression slightly higher than $\hat{\pi}$, A prefers that transgression to $\pi = \frac{k_D}{q}$, despite having to endure retaliation.

Proposition 1 *There is a unique subgame perfect Nash equilibrium to the transgression subgame, where $r^*(\pi)$ gives D's sequentially rational best-response and A's transgression choice is*

$$\pi^* = \begin{cases} \frac{k_D}{q} & \text{if } \overline{\pi} \in \left(\frac{k_D}{q}, \frac{k_D + qk_A}{q(1-q)}\right) \equiv \Delta(k_D, k_A, q) \\ \overline{\pi} & \text{otherwise.} \end{cases}$$

In canonical models (e.g., Levy 1988; Di Lonardo and Tyson 2022) transgressions are a coarse instrument, and assessing the effect of retaliation reduces to whether it "works" or "fails." Instead, in our framework deterrence manifests as an *effect* that measures by how much transgressions are reduced due to the threat of retaliation. To measure the impact of deterrence, the quantity of interest is the *deterrence effect*:

$$\overline{\pi} - \pi^*.$$

This highlights that deterrence is about a *counterfactual comparison* between the level of transgressions absent a credible retaliation threat by Defender, $\overline{\pi}$, and the level of transgressions in the presence of a credible retaliation threat, π^*. Proposition 1 identifies that the deterrence effect is characterized by an *extensive margin*, determining when the deterrence effect is present, and an *intensive margin*, determining how, when present, the deterrence effect changes with various parameters (in our model q, $\overline{\pi}$, and k_D).

Proposition 2 *The deterrence effect is strictly positive if and only if* $\overline{\pi} \in \Delta(k_D, k_A, q)$*, and zero otherwise. Moreover:*

1. ***Intensive margin****: when* $\overline{\pi} \in \Delta(k_D, k_A, q)$*, the deterrence effect is strictly increasing in* $\overline{\pi}$ *and* q*, and strictly decreasing in* k_D*;*

2. ***Extensive margin****: the set* $\Delta(k_D, k_A, q)$ *gets larger with increases in* k_D*,* k_A*, and* q*.*

Starting with the intensive margin, recall that when $\overline{\pi} \in \Delta(k_D, k_A, q)$ the level of transgressions pursued by A is $\frac{k_D}{q}$. Thus, the deterrence effect is strictly increasing in q and strictly decreasing in k_D. This is because at higher k_D, Defender is willing to tolerate higher transgressions without retaliating, therefore Aggressor is able to pursue greater transgressions without incurring retaliation. The former shows that when retaliation reduces the burden from a transgression on Defender by more (higher q), D becomes more willing to retaliate, increasing the deterrence effect.

The value of $\overline{\pi}$ is crucial to assess the importance of deterrence *to Defender*, whose incentives depend on what would materialize absent a retaliation threat. When the deterrence effect is positive, a (local) increase in $\overline{\pi}$ increases the deterrence effect because it increases the *counterfactual level of transgressions*, i.e., the transgressions that would have been pursued by A absent a retaliation threat.

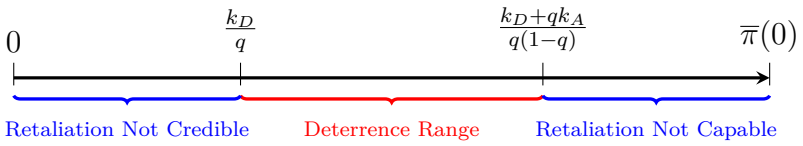


Figure 1: The Deterrence Range

The extensive margin is defined by the size of interval $\Delta(k_D, k_A, q) = \left(\frac{k_D}{q}, \frac{k_D+qk_A}{q(1-q)}\right)$, which

we call the *deterrence range*, and its "size" measures the extensive margin of deterrence

$$\frac{k_D + qk_A}{q(1-q)} - \frac{k_D}{q} = \frac{k_D + qk_A - (1-q)k_D}{q(1-q)} = \frac{k_D + k_A}{1-q} > 0.$$

The deterrence range is illustrated in Figure 1. Notice that it gets larger with increases in k_D, k_A, and q.[8] Increases in the deterrence range mean that transgressions that were initially too large to be deterred, or those that were initially too small to credibly retaliate against, *become deterrable*. That the deterrence range increases in k_A shows that when retaliation becomes more costly to endure for Aggressor, the threat of retaliation is more capable, and as a consequence, A is more willing to avoid retaliation and choose $\pi = \frac{k_D}{q}$.

An increase in q makes deterrence more credible, since it limits by a larger share the costs Defender endures following a transgression. However, while a larger q reduces the size of the maximum transgression that allows Aggressor to still avoid retaliation, it also means that a retaliation from Defender will reduce by more the benefits of a transgression for Aggressor, thus pushing Aggressor to settle for $\pi = \frac{k_D}{q}$. The overall effect is that the deterrence range increases in q.

Last, the deterrence range increases in k_D, meaning that as retaliation becomes more costly to execute for Defender, the deterrence range gets larger. This might be surprising given that an increase in k_D undermines the credibility of deterrence. As argued above, when retaliation becomes more costly, Defender is willing to tolerate a larger transgression without resorting to retaliation, i.e., $\frac{k_D}{q}$ increases. However, this also changes the calculus for Aggressor. Recall that Aggressor has a choice between the maximum possible transgression that does not trigger retaliation, $\frac{k_D}{q}$, and the maximum available transgression, $\overline{\pi}$, which does trigger retaliation. Consequently, increasing $\frac{k_D}{q}$ implies that Aggressor is able to transgress more without triggering retaliation. At the same time, the transgression that makes A

[8]Notice that since $\frac{k_D}{q} < \frac{k_D + qk_A}{q(1-q)}$, the deterrence range is nonempty.

indifferent between enduring retaliation and settling for $\frac{k_D}{q}$, $\hat{\pi}$ from (4), also goes up, and at a greater rate—because enduring retaliation means that A only enjoys $1-q$ of her transgression benefit. When taken together, although the lower bound of the deterrence range increases in k_D, so does the upper bound, $\frac{k_D+qk_A}{q(1-q)}$, and at a greater rate than the lower bound, thus making the deterrence range larger as k_D increases.

Empirical Implications for Deterrence

In approaching the empirical assessment of deterrence, previous work has highlighted how the influence of deterrent threats, when effective, is often hard to assess because of selection effects (Fearon 2002). This task is further complicated by the fact that, for a given set of parameters, a coercive tool might be able to deter some transgressions—those within the deterrence range—but not others. Therefore, measuring the effectiveness of coercive measures against a particular transgression requires consideration of how the bounds of the deterrence range are affected by relevant factors, namely studying the intensive and extensive margins of deterrence.

The intensive margin of deterrence reflects changes to the magnitude of the deterrence effect (when positive), and answers questions like "what factors enhance the effectiveness of coercive threats?" The extensive margin of deterrence determines the *prevalence* of cases for which the deterrence effect is positive (instead of 0). Consequently, changes that affect the extensive margin of the deterrence effect, by changing the scenarios that exhibit deterrence, introduce an "ecological effect" on the sample of cases (Glynn and Wakefield 2010). Together, the intensive and extensive margins of deterrence influence the *observed level of transgressions*, which is the level of transgressions when averaged over a sample of cases.

Empirical Implication 1 *The intensive and extensive margins of deterrence increase in q and k_A, and hence, the observed level of transgressions decreases in q and k_A.*

Since both the intensive and extensive margins increase in q, the effect of q on the observed level of transgressions is mutually reinforcing, and is strictly positive. An increase in q leads to a greater effectiveness of retaliatory threats and implies that retaliation accomplishes this effect in more cases. The cost retaliation imposes on Aggressor, k_A, has no effect on the intensive margin, but increases the extensive margin. Consequently, an increase in k_A does not reduce transgressions among countries where deterrence already does so, but it does increase the set of cases for which deterrence is an effective tool, and hence, reduces the observed level of transgressions.

Empirical Implication 2 *The extensive margin of deterrence increases, and the intensive margin decreases, with the cost retaliation imposes on Defender, k_D. Hence, increases in k_D have an ambiguous effect on the observed level of transgressions.*

The cost retaliation imposes on Defender, k_D, both increases the set of cases where deterrence works, and reduces the effect of deterrence in such cases. As a consequence, the effect of k_D on the observed level of transgressions is ambiguous. Specifically, consider how a shock that increases k_D would influence the observed level of transgressions. An increase in k_D will imply a reduction in the deterrence effect, $\overline{\pi} - \frac{k_D}{q}$, thus increasing the observed level of transgressions. At the same time, an increase in k_D will cause more countries to be deterred by D's threat of retaliation, which will be observed as an increase in countries choosing $\frac{k_D}{q}$ instead of $\overline{\pi}$. The overall effect on the observed level of transgressions, which averages over different cases, is ambiguous, and depends on which margin reflects a "steeper" change. Notice that this discussion does not imply that determining the overall influence of k_D is an empirical question, i.e., one cannot determine whether the extensive margin or intensive margin is more prevalent simply by measuring the observed level of transgressions Instead, one needs to identify both the extensive and intensive margins, and incorporate them into an estimation approach, or analyze them separately if possible.

Constraining and Coercing

In this section we consider constraining measures, which reduce the capacity of Aggressor to transgress against Defender's interests, focusing on how the presence of such tools influences Defender's decisions. To organize the subsequent analysis, we define Defender's optimal constraining choice as a function of a retaliation level, r. Specifically, define

$$x^*(r) \in \underset{x \geq 0}{\operatorname{argmin}}\ \overline{\pi}(x)(1 - rq) + rk_D + c(x). \tag{5}$$

This allows us to express a number of benchmarks as well as the equilibrium to our game.[9] In particular, by exogenously fixing $r = 0$, $x^*(0)$ gives the optimal level of constraining measures when retaliation is not possible for Defender, and similarly, by exogenously fixing $r = 1$, $x^*(1)$ gives Defender's optimal level of constraining measures when retaliation is going to be carried out regardless of the size of the transgression.

Lemma 2 *The optimal level of constraining measures is strictly decreasing in r. Moreover,*

(i) If retaliation is exogenously fixed at $r = 0$, then there is a unique subgame perfect equilibrium where Aggressor's choice of transgressions is $\pi^* = \overline{\pi}(x)$, and Defender's constraining choice is $x^*(0)$.

(ii) If retaliation is exogenously fixed at $r = 1$, then there is a unique subgame perfect equilibrium where Aggressor's choice of transgressions is $\pi^* = \overline{\pi}(x)$, and Defender's constraining choice is $x^*(1)$.

Lemma 2 shows that $x^*(0)$ is the maximum level of constraining measures that are pursued in any equilibrium, and that $x^*(1)$ is the minimum level of constraining measures that are consistent with equilibrium. Moreover, when the retaliation choice is unresponsive to the

[9] Note that because $\overline{\pi}''(x)(1 - rq) + c''(x) > 0$, $x^*(r)$ is unique and interior.

transgression level, the deterrence effect cannot arise since it is based on the contingency of retaliation.

Similar to the deterrence effect, the *constraining effect* is obtained by comparing what the level of transgressions would be with $x^*(0)$ to the transgression level absent constraining measures, $\overline{\pi}(0)$.[10] This counterfactual comparison is encapsulated in

$$\overline{\pi}(0) - \overline{\pi}(x^*(0)),$$

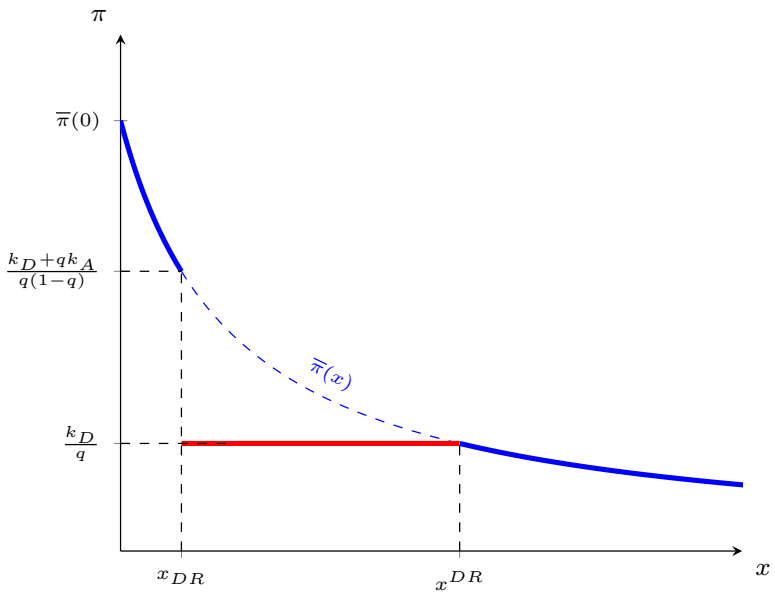
which is strictly positive and strictly increasing in pointwise increases in $\overline{\pi}$ and pointwise decreases in c. That is, the lower the cost of constraining, and the higher the costs of a transgression for Defender, the larger the constraining effect.

Figure 2 illustrates Defender's use of both diplomatic tools. On the horizontal axis we have the level of constraining measures and on the vertical axis is the level of transgressions. The downward sloping line reflects how constraining measures limit Aggressor's choices, since π cannot exceed this line. Starting from $\overline{\pi}(0)$, as x increases, the maximum transgression level A chooses decreases smoothly in x, until x_{DR}, at which point the threat of retaliation causes a discontinuous drop, because transgressions immediately above x_{DR} are not incentive compatible for Aggressor. This continues until x^{DR} and this is reflected by the dark line.

The analysis of the deterrence effect above corresponds to the last two stages of the full game, and thus, we can take the unique equilibrium to that subgame from Proposition 1 as given, noting that it depends on the level of constraining measures chosen in the first stage (substituting $\overline{\pi} = \overline{\pi}(x)$).

Proposition 3 *There exists a unique equilibrium,* $(x^*, \pi^*(x), r^*(\pi))$*, and* x^**, when strictly*

[10]Note that the benchmark level of constraining measures, $\overline{\pi}(0)$, connects to the benchmark level used to evaluate the deterrence effect by setting $\overline{\pi} = \overline{\pi}(0)$.



.

Figure 2: The Constraining Problem. The level x_{DR} is that which induces the bottom of the deterrence range and x^{DR} induces the highest level within the deterrence range.

positive, satisfies $x = x^*(r^*(\pi^*(x)))$, *otherwise, the unique equilibrium is* $x^* = 0$. *Moreover, there exists an* $\underline{x}$ *and an* $\overline{x}$, *with* $\underline{x} < \overline{x}$, *such that in equilibrium,* $x^* \notin (\underline{x}, \overline{x}]$.

This result, in addition to establishing that an equilibrium exists, also shows that there is a range of possible levels of constraining measures that are not consistent with equilibrium, $(\underline{x}, \overline{x}]$. In particular, because the threat of retaliation prevents a range of transgressions, there is a corresponding range, $(\underline{x}, \overline{x}]$, where the marginal benefit of higher constraining measures is zero. As a result, such intermediate levels of constraining measures are not pursued in any equilibrium.

The total amount of transgressions Defender is able to avoid, using constraining measures and retaliation threats, is assessed by considering the total effect from what Aggressor would

pursue, for a particular equilibrium $(x^*, \pi^*(x), r^*(\pi))$, which is written as

$$\overline{\pi}(0) - \pi^*(x^*),$$

and can be decomposed as

$$\overline{\pi}(0) - \pi^*(x^*) = \underbrace{\overline{\pi}(0) - \overline{\pi}(x^*)}_{\text{constraining effect}} + \overbrace{\overline{\pi}(x^*) - \pi^*(x^*)}^{\text{deterrence effect}}. \tag{6}$$

The total effect is the combination of the constraining effect, achieved using constraining measures, and the deterrence effect, achieved through retaliatory threats.

Recall that the initial maximum possible transgression available to Aggressor, $\overline{\pi}(0)$, reflects the strength of Aggressor relative to Defender, and we organize the remainder of our analysis around two salient cases. First, we call Aggressor *weak* if $\frac{k_D}{q} < \overline{\pi}(0) < \frac{k_D + qk_A}{q(1-q)}$, which says that Aggressor cannot (initially) pursue transgressions exceeding the deterrence range. Second, Aggressor is *strong* if $\overline{\pi}(0) > \frac{k_D + qk_A}{q(1-q)}$, which is when, absent constraining measures, Aggressor can pursue transgressions that would force Defender to retaliate (by Lemma 1). We ignore the case where $\overline{\pi}(0) < \frac{k_D}{q}$, because in this straightforward case, Aggressor is so weak that retaliation is never credible, and because our focus is on the relationship between constraining and coercive measures.

A Weak Aggressor

When Aggressor is weak, the maximum possible transgression that can be enacted does not yield high enough benefits to make it worth enduring retaliation by Defender. This is the case whenever $\overline{\pi}(0) \in \Delta(k_D, k_A, q)$, i.e., when the maximum transgression, absent constraining measures, is in the deterrence range.

Proposition 4 *Suppose Aggressor is weak, i.e.,* $\overline{\pi}(0) \in \Delta(k_D, k_A, q)$*, then the unique equilibrium level of constraining measures is*

$$x^* = \begin{cases} 0 & \text{if } \frac{k_D}{q} \leq \overline{\pi}(x^*(0)) + c(x^*(0)) \\ x^*(0) & \text{otherwise.} \end{cases}$$

This result shows that when Aggressor is weak, there are only two levels of constraining measures that are consistent with equilibrium. First, Defender pursues no constraints, in which case $x^* = 0$, and she relies solely on deterrence via the retaliation threat. Second, D chooses the same level of constraining measures that she would choose if retaliation were not available. Moreover, we see that as k_D gets large, or as q gets small, constraining measures become a more attractive tool for Defender.

Proposition 5 *When*

$$\frac{k_D}{q} \leq \overline{\pi}(x^*(0)) + c(x^*(0)), \tag{7}$$

the deterrence effect is strictly positive and the constraining effect is zero, whereas when (7) *fails, the deterrence effect is zero and the constraining effect is strictly positive.*

This result shows a conventional logic associated with how countries use, and potentially tradeoff, various tools in international affairs—they think of them as substitutes. In particular, Proposition 5 shows how introducing coercive tools causes Defender to either choose the same level of constraining measures, $x^*(0)$, or instead, to switch to relying solely on the threat of retaliation, i.e., $x^* = 0$. This can be seen in Figure 3. When constraining measures are a good tool at reducing the capacity of Aggressor, perhaps because they are uniquely vulnerable to asset seizures or decapitation strikes, then Defender uses them without relying on the threat of coercive tools. In this case,

$$\overline{\pi}(0) - \pi^*(x^*) = \overline{\pi}(0) - \overline{\pi}(x^*),$$

and the constraining effect is solely responsible for the reduction in transgressions. For constraining measures to be worth pursuing against a weak Aggressor, they have to be large enough to push the maximum possible transgression, $\overline{\pi}(x)$, outside of the deterrence range, at which point retaliation is not credible, and the deterrence effect is necessarily zero.

Instead, if constraining measures are relatively costly (or ineffective) for Defender, then the introduction of coercive tools leads her to abandon using them, and since $\pi^* = \frac{k_D}{q}$,

$$\overline{\pi}(0) - \pi^*(x) \mid_{x^*=0} = \overline{\pi}(0) - \frac{k_D}{q},$$

implying that the retaliation threat reduces transgressions solely via the deterrence effect.

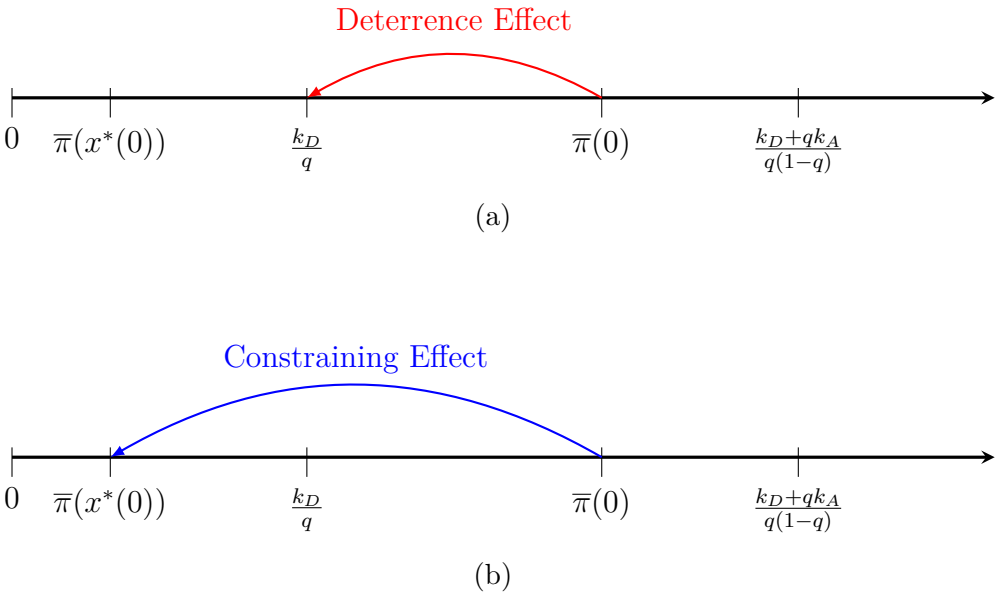


Figure 3: Equilibrium transgression choice: (a) Constraining measures are too costly relatively to their effectiveness, and the threat of retaliation is the only coercive tool that limits transgressions; (b) Constraining measures are effective enough relative to their costs, Defender imposes a large enough constraints that make the threat of retaliation not credible.

A Strong Aggressor

We now consider when Aggressor is strong, i.e., $\overline{\pi}(0) > \frac{k_D+qk_A}{q(1-q)}$. This means that, absent constraining measures, Aggressor is capable of seizing large enough benefits through transgressions to compensate for enduring retaliation. Importantly, a strong Aggressor presents a challenging scenario for Defender, due to the fact that the threat of coercive intervention is not capable (despite it being credible), and thus the threat of retaliation alone will not induce Aggressor to moderate the extent of the transgression.

Lemma 3 *When Aggressor is strong, $\overline{\pi}(0) > \frac{k_D+qk_A}{q(1-q)}$, in any equilibrium, Defender chooses a positive level of constraining measures, i.e., $x^* > 0$, and the level of transgressions, π^*, is strictly lower than if only retaliation were available.*

This result establishes that constraining measures are always part of how a Defender responds to a strong Aggressor. Because Aggressor is strong, degrading their capacity to transgress is always optimal, and the question is about the extent of constraining measures that are used. Moreover, when compared to the counterfactual world where the possibility of constraining measures are absent, and retaliation is the only tool available to Defender, transgressions become smaller. Without constraining measures, A's transgression is going to be equal to $\overline{\pi}(0)$, and since $\overline{\pi}(0)$ exceeds the deterrence range, Defender is forced to retaliate. This simple observation shows that constraining measures, and retaliatory threats, are not substitutes when Aggressor is strong.

Proposition 6 *Suppose Aggressor is strong, i.e., $\overline{\pi}(0) > \frac{k_D+qk_A}{q(1-q)}$, then the unique equilibrium level of constraining measures, x^*, can take three values: $x^*(1)$, $x^*(0)$, or $x^\dagger$, defined by*

$$x^\dagger = \overline{\pi}^{-1}\left(\frac{k_D + qk_A}{q(1-q)}\right),$$

and where the equilibrium values satisfy $x^*(1) < \underline{x} < x^\dagger < x^*(0)$.

The interaction between constraining measures and the threat of retaliation is more subtle when Aggressor is strong. Proposition 6 establishes that there are three different levels of constraining measures, $x^*(1)$, $x^\dagger$, and $x^*(0)$. Which of these three is the unique equilibrium level of constraining measures depends on the functions, $\overline{\pi}(\cdot)$ and $c(\cdot)$, which reflect the effectiveness of constraining measures to curb Aggressor's transgression capabilities, and the cost of implementing constraining measures, respectively. First, consider when the unique equilibrium level of constraining measures are $x^*(1)$ or $x^*(0)$.

Proposition 7 *Suppose Aggressor is strong, i.e.,* $\overline{\pi}(0) > \frac{k_D+qk_A}{q(1-q)}$*, and that the unique equilibrium level of constraining measures,* x^**, is* $x^*(1)$ *or* $x^*(0)$*, then the deterrence effect is zero and the constraining effect is strictly positive.*

Proposition 7 focuses on two of the three possible equilibria that can emerge when Aggressor is strong, which are illustrated in Figure 4. The first is when constraining measures, represented by $\overline{\pi}(\cdot)$, are highly effective relative to their cost, $c(\cdot)$. In this case the unique equilibrium level of constraining measures are $x^*(0)$, and they are used at such a high level that they constrain what Aggressor can pursue so much that retaliation lacks credibility (i.e., (3) fails). Since retaliation is not credible, the deterrence effect is zero.

The second equilibrium level of constraining measures, $x^*(1)$, by contrast, is where constraining measures are ineffective relative to their cost, $c(\cdot)$. In this case, Defender chooses such a low level of constraining measures, that retaliation lacks capability, and cannot motivate A to reduce transgressions. It is important to note that when $x^*(1)$ is the equilibrium level of constraining measures, they are observed in conjunction with the actual use of retaliation, which manifests either as economic sanctions or military force. In this case, retaliation reduces transgressions by $1 - q$.

With a strong Aggressor, constraining measures and coercive tools, such as punitive

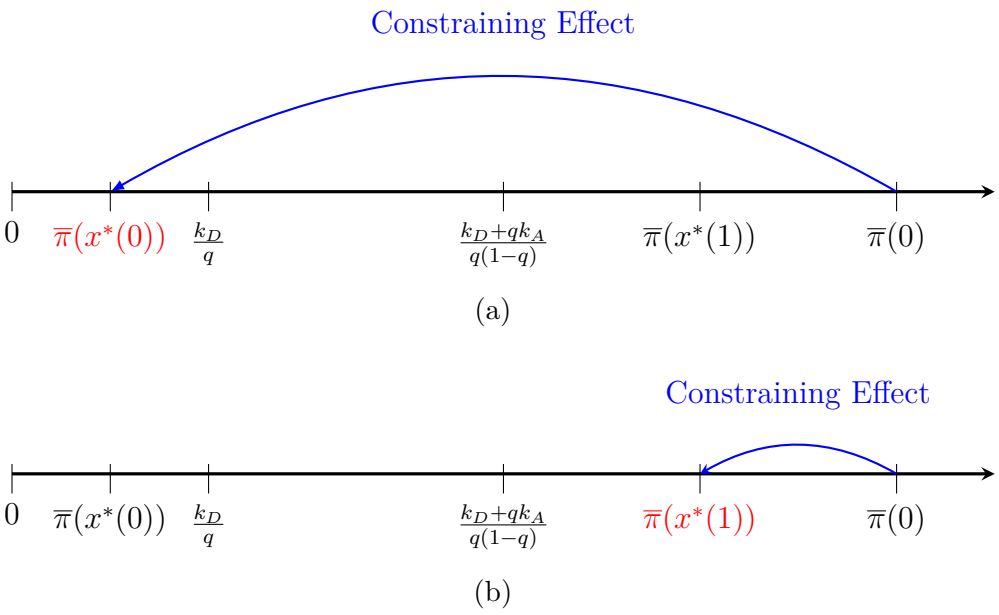


Figure 4: Equilibrium transgression choice: (a) Defender relies uniquely on constraining measures even when retaliation is an option; (b) Defender imposes limited constraints on Aggressor, and retaliates.

forms of economic sanctions or military actions, interact in a way that they do not in the case of a weak Aggressor. Proposition 7 considers when the effectiveness of constraining measures, relative to their cost, was either high or low. But what if constraining measures are relatively cheap at low levels but become extremely costly at high levels? There are many examples where this is case, such as targeted freezing of assets, which prevent arms sales that would otherwise take place while leaving routine economic activity largely unaffected, but as financial constraints are ratcheted up to include more pedestrian transactions, this can have a large impact on economic activity in Defender, and may not be worth the political cost. As a consequence, and different from the case of a weak Aggressor, constraining measures and the threat of retaliation *work together*.

Proposition 8 *Suppose Aggressor is strong, i.e.,* $\overline{\pi}(0) > \frac{k_D+qk_A}{q(1-q)}$*, and that the unique equilibrium level of constraining measures is* $x^* = x^\dagger$*, then the deterrence effect and the constraining*

effect are strictly positive.

This result highlights when constraining measures and coercive threats are used together. When the unique equilibrium level of constraining measures is $x^\dagger$, we can decompose the total effect given generally in (6), where $x^* = x^\dagger$, and since $\pi^* = \frac{k_D}{q}$, as

$$\overline{\pi}(0) - \pi^*(x^\dagger) = \underbrace{\overline{\pi}(0) - \overline{\pi}(x^\dagger)}_{\text{constraining effect}} + \overbrace{\overline{\pi}(x^\dagger) - \frac{k_D}{q}}^{\text{deterrence effect}}.$$

Since both the constraining effect and deterrence effect are strictly positive, we see how constraining and coercive tools are used in conjunction with each other—as complements—because they enhance each others' effectiveness at reducing transgressions from Aggressor. Moreover, Defender is better off than using either of these tools in isolation.[11]

When the unique equilibrium level of constraining measures is $x^\dagger$, the maximum transgression Aggressor can pursue is reduced to $\overline{\pi}(x^\dagger) = \frac{k_D + qk_A}{q(1-q)}$, which is the upper bound of the deterrence range. At this point retaliation is both credible and capable, and as a result, the transgression A chooses further reduces from $\overline{\pi}(s^\dagger)$ to $\frac{k_D}{q}$—as a result of the deterrence effect. We call this *constraining to deter*, because it reflects the use of constraining measures as a way to grant capability to retaliatory threats, and it is illustrated in Figure 6. Defender constrains Aggressor's capability and *activates* the deterrence effect.

To see what the introduction of retaliation accomplishes, we contrast the equilibrium level of constraining measures when they equal $x^\dagger$, to what D would choose if retaliation were not available, i.e., the (in this case) out-of-equilibrium $x^*(0)$. By Proposition 2, $x^*(0)$ is out-of-equilibrium whenever it falls below $\overline{x}$. When this is true, the level of constraining measures

[11]The optimal level of constraining measures chosen when retaliation is not an option, $x^*(0)$, is still available for Defender, thus in case Defender opts for a different level once retaliation is an option, it implies that this different level of constraining measures ensures a higher payoff for Defender.

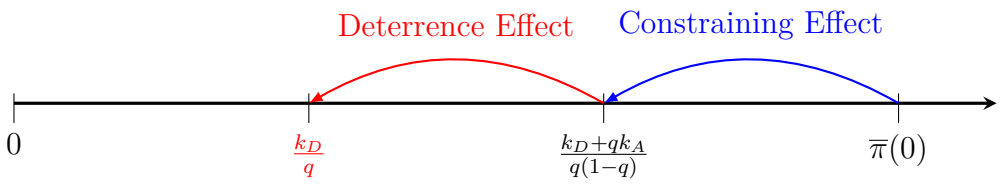


Figure 5

Figure 6: Constraining to Deter

has increased when compared to the counterfactual where retaliation is not available. As a result, *the introduction of retaliation increases the use of constraining measures.* This feature is another way to view constraining to deter, as it shows how constraining measures and retaliation threats work together, and thus serve as complementary instruments in Defender's diplomatic toolkit.

Empirical Implications

We conclude by discussing some of the implications that our results have for the empirical study of diplomacy. Recall that our model and results stress the importance of distinguishing different tools of diplomacy by their strategic role, rather than along descriptive dimensions, such as whether force is used. Specifically, it is important not to conflate military actions with coercive actions, because one might erroneously conclude that observing military actions is a sign that deterrence has failed completely (Lebow and Stein 1989, 1990).

We organize our discussion around two themes. First, assessing the effectiveness of various tools of diplomacy, whether used coercively or as constraining measures, requires careful assessment of an appropriate counterfactual. Second, similar to our treatment of deterrence above, determining the impact of different diplomatic tools involves understanding the sample composition of conflict dyads. This in turn, requires identifying the deterrence range, since this determines when an aggressor is weak instead of strong, and how constraining and

coercive tools are used.

Assessing the total effectiveness of constraining measures or coercive threats requires knowing the counterfactual level of transgressions that would be pursued in the absence of such tools.

Empirical Implication 3 *The level of transgressions absent constraining measures or coercive threats, $\overline{\pi}(0)$, is never observed, and hence, the total effectiveness of constraining measures or coercive threats cannot generally be assessed.*

Since the total effectiveness of diplomatic tools requires evaluating the counterfactual comparison: $\overline{\pi}(0) - \pi^*(x^*)$. Because $\overline{\pi}(0)$ is never achieved as an equilibrium outcome in our game, the total effectiveness of constraining measures or coercive threats remains latent. Consequently, measuring this latent quantity will require a shock that shuts down (at random) the opportunity for Defender to use either constraining or coercive measures, i.e., all the strategic channels isolated by our model. Thus, a key empirical implication that follows from our model is that questions about the effectiveness of constraining or coercive tools are likely out of the reach of most empirical exercises.

Restricting to the evaluation of constraining measures, we now consider what kinds of inferences can be drawn from comparisons between cases where constraining measures are used, $x^* > 0$, to cases where they are not, $x^* = 0$, i.e., the comparison:

$$\pi^*(x^*) \Big|_{x^*>0} - \frac{k_D}{q}. \tag{8}$$

We begin by considering when such a comparison is relatively straightforward before considering complications.

Empirical Implication 4 *Restricting to a sample of weak Aggressor-Defender dyads, a comparison of transgressions when $x^* = 0$ to when $x^* > 0$ provides a biased estimate of the total effectiveness of constraining measures.*

In the case of weak Aggressor-Defender, taking the level of transgressions when $x^* = 0$ as representing the counterfactual level $\overline{\pi}(0)$ would be misleading because it severely underestimates the total constraining effect. This is because when $x^* = 0$, the weak Aggressor chooses $\pi^* = \frac{k_D}{q} < \overline{\pi}(0)$, and hence, the observed difference from (8) is $\frac{k_D}{q} - \overline{\pi}(x^*(0))$ instead of $\overline{\pi}(0) - \overline{\pi}(x^*(0))$. We can write (8) as

$$\frac{k_D}{q} - \overline{\pi}(x^*(0)) = \overbrace{\overline{\pi}(0) - \overline{\pi}(x^*(0))}^{\text{True effect}} + \underbrace{\frac{k_D}{q} - \overline{\pi}(0)}_{\text{bias}},$$

showing that the (unobserved) deterrence effect creates bias when assessing the total effectiveness of constraining measures. The threat of retaliation essentially constitutes an omitted variable. Consequently, one can only evaluate the effectiveness of constraining measures *relative to the credible threat of retaliation* rather than their overall effectiveness.

Empirical Implication 4 follows from the restriction to weak Aggressor-Defender dyads. However, the deterrence range in a particular scenario (or dyad) depends on the specific relationship between Aggressor and Defender. Consequently, factors such as the military balance between Aggressor and Defender, as well as the importance of the issue in contention (i.e., the status quo), affect the deterrence range. Together, these factors determine whether an individual dyad is one of a weak or a strong Aggressor. The deterrence range is not specific to a particular Defender, but it is specific to a particular Aggressor-Defender dyad. Specifically, a single aggressor could be weak vis-a-vis one Defender, but strong vis-a-vis another. Consequently, identifying whether a particular dyad is one with a weak Aggressor, or one with a strong Aggressor, is not generally possible case by case.

Proposition 6 implies that there are three potential levels of constraining measures $x^*(0)$, $x^*(1)$, and $x^\dagger$ when Aggressor is strong. Without a sharp distinction between weak vs strong Aggressor-Defender dyads, comparison (8) is less straightforward, and illustrates some novel issues. Suppose, for instance, that in dyad i, $x_i^* > 0$, and in dyad j, $x_j^* = 0$, then the

empirical comparison closest to (8) becomes

$$\pi_i^*(x_i^*) - \frac{k_D^j}{q^j}, \tag{9}$$

which we can decompose into

$$\overbrace{\pi_i^*(x_i^*) - \frac{k_D^i}{q^i}}^{\text{Comparison (8)}} + \underbrace{\frac{k_D^i}{q^i} - \frac{k_D^j}{q^j}}_{\text{baseline difference}}.$$

The first term reflects the substantive comparison for dyad i, giving the level of transgressions achieved through coercive threats from those achieved when constraining measures are used. The second term captures the baseline difference between dyad i and dyad j because the level of transgressions below which retaliation lacks credibility is not the same across the two cases. This matters, for instance, if dyad i is a case of constraining to deter, where the level of transgressions is $\frac{k_D^i}{q^i}$, implying that the first term is zero, and hence, (9) would reflect only baseline differences across cases rather than any substantively important difference.[12] This analysis applies also to the cases when $x_i^* = x_i^*(0)$ and $x_i^* = x_i^*(1)$, and becomes more complicated if the analyst cannot distinguish the three equilibria of Proposition 6.

The above argument illustrates the importance of distinguishing, at least on average, cases where the aggressor is weak from those where she is strong. Moreover, it also shows that baseline differences across cases can introduce bias in natural comparisons such as (8). One possible solution would be to use a measure of military strength to proxy for whether an Aggressor-Defender dyad is (or is more likely to be) one with a weak v strong Aggressor. Such an approach would need a source of exogenous variation that alters the average cutoff dividing weak from strong aggressors, similar to a fuzzy regression discontinuity design.

[12] Specifically, $\pi_i^*(x_i^*) - \frac{k_D^j}{q^j} = \frac{k_D^i}{q^i} - \frac{k_D^j}{q^j}$.

Importantly, any valid instrument could not also alter the effects of interest, suggesting that it could not influence the cutoff through k_D or q, since this would influence the level of transgressions. In sum, our discussion illustrates that the empirical evaluation of diplomacy is a somewhat involved exercise, which leverages both a theoretical understanding of the composition of samples, as well as sophisticated methodological approaches that can be applied only after the strategic composition of the sample is well understood.

Conclusion

Our analysis clarifies how constraining measures, used prior to the initiation of an international crisis, and coercive deterrent threats, using either punitive economic sanctions or military actions, work together as part of a country's diplomatic toolkit. To study the role of different kinds of constraining measures or retaliatory threats, we build on canonical models of deterrence, departing on two key dimensions. First, we allow the Aggressor to choose how much of a transgression to pursue, e.g., how much they want to alter the status quo. This leads to a more nuanced formulation of deterrence because the threat of retaliation in our framework operates as an "effect" which compares transgressions with and without potential retaliation. Second, we introduce the possibility of constraining measures that shape the international crisis by preemptively reducing how much Aggressor can transgress.

We present a number of results, which when taken together, highlight the importance of Aggressor's initial strength relative to Defender. When Aggressor is weak, then Defender uses either constraining measures or deterrence but not both, which is determined by how effective actions that constrain Aggressor's opportunities, relative to the costs they impose on Defender. Instead, when Aggressor is strong, then constraining measures and deterrence work together to constrain Aggressor's transgressions. Our main result when Aggressor is strong is to isolate *constraining to deter*, which identifies when Defender uses constraining

measures preemptively to activate deterrent threats that would not be effective otherwise.

Going back to our opening example, Ukrainian President Volodymyr Zelenskiy understood the use of economic sanctions as a preemptive measure that would alter the military capability/resolve of the Russian military, whereas US Secretary of State Antony Blinken understood economic sanctions as a coercive, or punitive, measure. We show that economic sanctions, as well as military actions, can function as both a preemptive measure that constrains as well as a coercive deterrent threat. Which function a particular tool serves depends on their strategic role and not whether a particular action is administered by the military or a financial/economic body. We also explore complications arising when assessing what role a particular tool serves, arguing that having a clear idea what function a particular action serves in a broader strategic context is critical.

Appendix

Proof of Lemma 1: In the text. ■

Proof of Proposition 1: Follows from Lemma 1 and the discussion in the text. ■

Proof of Proposition 2: Follows from the discussion after the statement in the text. ■

Proof of Lemma 2: Let $r' > r$, then take the difference

$$\begin{aligned}\overline{\pi}(x)(1-r'q) + r'k_D + c(x) - (\overline{\pi}(x)(1-rq) + rk_D + c(x)) \\ = (r-r')(q\overline{\pi}(x) - k_D).\end{aligned}$$

Differentiating this difference by x yields

$$(r-r')q\overline{\pi}'(x),$$

which since $r' > r$ and $\overline{\pi}'(x) < 0$ for all x, implies that Defender's payoff, (1), has strict increasing differences between r and x. Thus, the strict Monotonicity Theorem I of Edlin and Shannon (1998), applied to minimization problems, establishes that $x^*(r)$ is strictly decreasing in r. ■

Proof of Proposition 3: Taking $r^*(\pi)$ and $\pi^*(x)$ from Proposition 1, and $x^*(r)$ from Lemma 2. For an equilibrium, the level of constraining measures must be a sequential best-response to the downstream retaliation and transgression decisions, each of which depend on x (directly or indirectly). There are two possibilities. First, from (5), any level x that satisfies

$$x = x^*(r^*(\pi^*(x))), \tag{10}$$

is an equilibrium level of constraining measures, x^*. Since the left-hand side is strictly increasing in x, and the right-hand side is weakly decreasing in x, the equilibrium is unique.

Second, if transgression level $\frac{k_D}{q}$ leaves D better off than a strictly higher x that satisfies (10), then $x^* = 0$ is the unique equilibrium.

For the second part, using the deterrence range, define three levels of constraining measures as follows. First, $\overline{x}$ is the value of x that solves

$$\overline{\pi}(x) + c(x) = \frac{k_D}{q}.$$

The set $(\overline{\pi}^{-1}(\frac{k_D}{q}), \overline{x})$ is the set of constraining measures for which D strictly prefers $\frac{k_D}{q}$ to the transgression level that would be achieved through such constraining measures. Second, $\hat{x}$ is the value of constraining measures, x, that solves

$$(1-q)\overline{\pi}(x) + k_D + c(x) = \frac{k_D}{q}.$$

Third, define $x^\dagger = \overline{\pi}^{-1}\left(\frac{k_D + qk_A}{q(1-q)}\right)$, and then the value of constraining measures:

$$\underline{x} = \min\{\hat{x}, x^\dagger\}.$$

The set $(\hat{x}, x^\dagger)$ is the set of constraining measures for which D would strictly prefer $\hat{x}$ to $x^\dagger$, when it is nonempty.

To show the last part, suppose, by contradiction, that there exists an equilibrium in which Defender chooses $x' \in (\underline{x}, \overline{x}]$. Then, since $\overline{\pi}(x') \in \Delta(k_D, k_A, q)$, in the subgame starting with Aggressor's transgression choice, by Proposition 1, Aggressor chooses $\pi^* = \frac{k_D}{q}$. Also by Proposition 1, if Defender had chosen $x'' < x'$, then Aggressor would choose $\pi^* = \frac{k_D}{q}$. Since c is strictly increasing, Defender is then better off choosing x'' instead of x', contradicting that x' was part of an equilibrium strategy profile. Since x' was arbitrary, this establishes the claim. ■

Proof of Proposition 4: Since A is weak, $\overline{\pi}(0) \in \Delta(k_D, k_A, q)$. If constraining measures

are strictly positive then, by Proposition 3, they must exceed $\underline{x}$, in which case retaliation is not credible, and hence $r^* = 0$. Then, from (10), we have that $x = x^*(0)$, establishing that the equilibrium level of constraining measures must be $x^*(0)$. Instead, if constraining measures are $x = 0$, then by Proposition 1, $\pi^*(0) = \frac{k_D}{q}$, and then $r^*(\frac{k_D}{q}) = 0$, thus by (10), $x^* = 0$. The latter is the unique equilibrium level of constraining measures if and only if

$$\frac{k_D}{q} \leq \overline{\pi}(x^*(0)) + c(x^*(0)),$$

and otherwise, $x^*(0)$ is the unique equilibrium level of constraining measures. ■

Proof of Proposition 5: Follows from Proposition 4 and the discussion in the text. ■

Proof of Lemma 3: Suppose that Aggressor is strong, so that $\overline{\pi}(0) > \frac{k_D + qk_A}{q(1-q)}$, and there is an equilibrium in which $x^* = 0$. Since, by optimality,

$$-(1-q)\overline{\pi}(x^*(1)) - k_D - c(x^*(1)) \geq -(1-q)\overline{\pi}(0) - k_D, \tag{11}$$

it is sufficient to show that $x^*(1) > 0$. Inequality (11) reduces to

$$\frac{\overline{\pi}(0) - \overline{\pi}(x^*(1))}{c(x^*(1))} \geq \frac{1}{1-q}.$$

Taking $x^*(1) \to 0$, and applying L'Hôpital's rule, yields

$$-\lim_{x^*(1)\to 0} \frac{\overline{\pi}'(x^*(1))}{c'(x^*(1))} \geq \frac{1}{1-q},$$

which becomes strict since $\lim_{x\to 0} c'(x) = +\infty$, implying that $x^*(1) > 0$. ■

Proof of Proposition 6: Since A is strong, $\overline{\pi}(0) > \frac{k_D + qk_A}{q(1-q)}$. By Lemma 3, constraining measures are strictly positive, so by Proposition 3, they must exceed $\overline{x}$ or must satisfy $x^* \leq \underline{x}$. In the former, retaliation is not credible, and hence $r^* = 0$, in which case, from (10), the

equilibrium level of constraining measures must be $x^*(0)$. If $x^* \leq \underline{x}$, there are two relevant cases. First, if $x^* < \underline{x}$, then retaliation is credible but not capable, and hence, D retaliates on the path of play. In this case, $\pi^* = \overline{\pi}(x)$ and $r^*(\overline{\pi}(x)) = 1$, and hence, from (10), the equilibrium level of constraining measures must be $x^* = x^*(1)$. Second, if $x^* = \underline{x}$, then by Proposition 1, $\pi^* = \frac{k_D}{q}$, and then $r^* = 0$. The latter yields $x^* = x^\dagger$ and is the unique equilibrium level of constraining measures if and only if

$$\frac{k_D}{q} + c(\overline{x}) \leq \min\{(1-q)\overline{\pi}(x^*(1)) + k_D + c(x^*(1)), \overline{\pi}(x^*(0)) + c(x^*(0))\}.$$

Moreover, $x^*(0)$ is the unique equilibrium level of constraining measures when

$$\overline{\pi}(x^*(0)) + c(x^*(0)) \leq \min\{(1-q)\overline{\pi}(x^*(1)) + k_D + c(x^*(1)), \frac{k_D}{q} + c(\overline{x})\},$$

with $x^*(1)$ being the unique equilibrium level of constraining measures when

$$(1-q)\overline{\pi}(x^*(1)) + k_D + c(x^*(1)) \leq \min\{\overline{\pi}(x^*(0)) + c(x^*(0)), \frac{k_D}{q} + c(\overline{x})\}.$$

■

Proof of Proposition 7: Follows by discussion in the text. ■

Proof of Proposition 8: Follows by discussion in the text. ■

References

Baliga, Sandeep, Ethan Bueno de Mesquita and Alexander Wolitzky. 2020. "Deterrence with imperfect attribution." *American Political Science Review* 114(4):1155–1178.

Baliga, Sandeep and Tomas Sjöström. 2008. "Strategic ambiguity and arms proliferation." *Journal of political Economy* 116(6):1023–1057.

Baliga, Sandeep and Tomas Sjöström. 2022. "Optimal Compellence." *Working Paper* .

Bapat, Navin A and Bo Ram Kwon. 2015. "When are sanctions effective? A bargaining and enforcement framework." *International Organization* 69(1):131–162.

Bergman, Ronen. 2018. *Rise and Kill First: The Secret History of Israel's Targeted Assassinations*. Hachette UK.

Bils, Peter and Bradley C Smith. 2023. "The Logic of Secret Alliances." *Working Paper* .

Chassang, Sylvain and Gerard Padró i Miquel. 2010. "Conflict and deterrence under strategic risk." *The Quarterly Journal of Economics* 125(4):1821–1858.

Clark, David H and William Reed. 2005. "The strategic sources of foreign policy substitution." *American Journal of Political Science* 49(3):609–624.

Di Lonardo, Livio and Scott A. Tyson. 2022. "Political Instability and the Failure of Deterrence." *The Journal of Politics* 84(1):180–193.

Edlin, Aaron S and Chris Shannon. 1998. "Strict monotonicity in comparative statics." *Journal of Economic Theory* 81(1):201–219.

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

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

Fearon, James D. 2002. "Selection effects and deterrence." *International Interactions* 28(1):5–29.

Giumelli, Francesco. 2011. *Coercing, constraining and signalling: explaining UN and EU sanctions after the Cold War*. ECPR press.

Glynn, Adam N and Jon Wakefield. 2010. "Ecological inference in the social sciences." *Statistical methodology* 7(3):307–322.

Grauvogel, Julia, Amanda A Licht and Christian von Soest. 2017. "Sanctions and signals: How international sanction threats trigger domestic protest in targeted regimes." *International Studies Quarterly* 61(1):86–97.

Grauvogel, Julia, Nikolay Marinov and Tsz-Ning Wong. 2022. "Targeted Sanctions Against Authoritarian Elites." *Working Paper* .

Grillo, Edoardo and Antonio Nicolò. 2023. "Learning it the hard way: Conflicts, economic sanctions and military aids.".

Gurantz, Ron and Alexander V Hirsch. 2017. "Fear, appeasement, and the effectiveness of deterrence." *The Journal of Politics* 79(3):000–000.

Hörner, Johannes, Massimo Morelli and Francesco Squintani. 2015. "Mediation and peace." *The Review of Economic Studies* 82(4):1483–1501.

Jackson, Matthew O and Massimo Morelli. 2011. The reasons for wars: an updated survey. In *The handbook on the political economy of war*. Edward Elgar Publishing.

Joseph, Michael F. 2023. "Do Different Coercive Strategies Help or Hurt Deterrence?" *International Studies Quarterly* 67(2):sqad018.

Kavaklı, Kerim Can, J Tyson Chatagnier and Emre Hatipoğlu. 2020. "The power to hurt and the effectiveness of international sanctions." *The Journal of Politics* 82(3):879–894.

Kirshner, Jonathan. 1997*a*. *Currency and coercion: the political economy of international monetary power*. Princeton University Press.

Kirshner, Jonathan. 1997*b*. "The microfoundations of economic sanctions." *Security Studies* 6(3):32–64.

Kustra, Tyler. 2023. "Economic Sanctions as Deterrents and Constraints." *Journal of Peace Research* 60(4):649–660.

Kydd, Andrew H and Roseanne W McManus. 2017. "Threats and assurances in crisis bargaining." *Journal of conflict resolution* 61(2):325–348.

Lebow, Richard Ned and Janice Gross Stein. 1989. "Rational deterrence theory: I think, therefore I deter." *World politics* 41(2):208–224.

Lebow, Richard Ned and Janice Gross Stein. 1990. "Deterrence: The elusive dependent variable." *World politics* 42(3):336–369.

Lektzian, David J and Christopher M Sprecher. 2007. "Sanctions, signals, and militarized conflict." *American Journal of Political Science* 51(2):415–431.

Levy, Jack S. 1988. "When do deterrent threats work?" *British Journal of Political Science* 18(4):485–512.

Marinov, Nikolay. 2005. "Do economic sanctions destabilize country leaders?" *American Journal of Political Science* 49(3):564–576.

McCormack, Daniel and Henry Pascoe. 2017. "Sanctions and preventive war." *Journal of Conflict Resolution* 61(8):1711–1739.

McLean, Elena V and Taehee Whang. 2010. "Friends or foes? Major trading partners and the success of economic sanctions." *International Studies Quarterly* 54(2):427–447.

Meirowitz, Adam, Massimo Morelli, Kristopher W Ramsay and Francesco Squintani. 2019. "Dispute resolution institutions and strategic militarization." *Journal of Political Economy* 127(1):378–418.

Paine, Jack and Scott A Tyson. 2020. The Uses and Abuses of Formal Models in Political Science. In *SAGE Handbook of Political Science: A Global Perspective*, ed. Dirk Berg-Schlosser, Bertrand Badie and Leonardo Morlino. Sage Publications Sage CA: Thousand Oaks, CA.

Pape, Robert A. 1997. "Why economic sanctions do not work." *International security* 22(2):90–136.

Powell, Robert. 1987. "Crisis bargaining, escalation, and MAD." *American Political Science Review* 81(3):717–735.

Powell, Robert. 1989. "Nuclear deterrence and the strategy of limited retaliation." *American Political Science Review* 83(2):503–519.

Powell, Robert. 1990. *Nuclear deterrence theory: The search for credibility*. Cambridge University Press.

Ramsay, Kristopher W. 2017. "Information, uncertainty, and war." *Annual Review of Political Science* 20:505–527.

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

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Harvard university press.

Schram, Peter. 2021. "Hassling: how states prevent a preventive war." *American journal of political science* 65(2):294–308.

Schram, Peter. 2022. "When Capabilities Backfire: How Improved Hassling Capabilities Produce Worse Outcomes." *The Journal of Politics* 84(4):2246–2260.

Schultz, Kenneth A. 1998. "Domestic opposition and signaling in international crises." *American Political Science Review* 92(4):829–844.

Slantchev, Branislav L. 2011. *Military threats: the costs of coercion and the price of peace*. Cambridge University Press.

Smith, Bradley C and William Spaniel. 2019. "Militarized disputes, uncertainty, and leader tenure." *Journal of Conflict Resolution* 63(5):1222–1252.

Spaniel, William. 2021. "Bargaining over Costly Signals." *Working Paper* .

Spaniel, William and Bradley C Smith. 2015. "Sanctions, uncertainty, and leader tenure." *International Studies Quarterly* 59(4):735–749.

Spaniel, William and Iris Malone. 2019. "The uncertainty trade-off: Reexamining opportunity costs and war." *International Studies Quarterly* 63(4):1025–1034.

Spaniel, William and Işıl İdrisoğlu. 2023. "Endogenous military strategy and crisis bargaining." *Conflict Management and Peace Science* Forthcoming.

Wagner, R Harrison. 1992. "Rationality and misperception in deterrence theory." *Journal of Theoretical Politics* 4(2):115–141.

Wolford, Scott. 2007. "The turnover trap: New leaders, reputation, and international conflict." *American Journal of Political Science* 51(4):772–788.

Wolford, Scott. 2014. "Showing restraint, signaling resolve: Coalitions, cooperation, and crisis bargaining." *American Journal of Political Science* 58(1):144–156.