Rationalism, Social Constructivism and Coercion *

Steven Beard†
Post-doctoral Researcher
Center for Peace and Security Studies
University of California, San Diego

July 15, 2020

Abstract

Coercion, where one actor threatens to impose costs to get a second actor to do or refrain from doing something, is a central component of international relations. Nuclear threats, coercive bombing, economic sanctions, trade disputes, terrorism, and guerrilla warfare all represent coercive attempts. A long line of research has examined coercion, especially regarding how punishment threats can be used to deter adversaries. In this paper, I reexamine the basic logic of when coercion is successful using a series of formal models. I show that previous research has missed a couple key elements in understanding coercive success. First, explanations for coercive success are different depending on the actors' time horizons. When the actors' time horizons are short, coercive success is determined primarily by the existence and form of available commitment devices, such as the ability to tie hands or incur audience costs. However, when time horizons are long, there are generally multiple equilibria and the effects of reputation override the importance of commitment devices. My second novel finding that with long time horizons, coercive success is determined primarily by socially constructed intersubjective beliefs, as this is necessary to determine which equilibrium actually occurs. Thus, even hard power may depend on socially constructed elements.

*I appreciate the comments of Christopher Butler, Megan Shannon, David Bearce, Shoko Kohama, Karl Sorenson, Erik Gartzke, Michael Rubin, Matt Millard, David Kang, (among others)

†sbeard@ucsd.edu

1 Introduction

Coercion through punishment is a key element in international relations and political science more broadly. The actual definition of coercion is somewhat vague. In this paper, I will refer specifically to cases where the sender of the coercive threat threatens to punish or impose costs on the target of coercion if the target does or fails to do what the sender wants. Notably, in this scenario the threatened punishment does not directly achieve any of the sender's goals. Thus, to achieve anything, the sender is relying on the target voluntarily complying under the threat of punishment if they don't. Among other things, sanctions, coercive bombing, nuclear retaliation, retaliatory tariffs, terrorism, and guerrilla war would all appear to represent types of coercion in international relations. Beyond IR, some intergovernmental bargaining might take the form of coercion, as might political protests. In fact, legal punishment regimes are often a form of coercion - threatening punishment through fines or prison to ensure compliance with laws, and thus the very existence of the state is tied up with coercion.

In this paper, I reevaluate when coercion is successful using a series of formal models. The first model is a basic, model of coercion. In this model, the target chooses some division of a disputed good. As the sender cannot directly achieve their goals, this division is implemented regardless of the sender's actions. After the target has divided the good, the sender can choose to punish the target, which would also be costly to the sender, or not. In this model, coercion never works as it would be irrational for the sender to actually carry out the punishment threat, assuming punishment is mutually costly.

Therefore, I examine a variation on this basic model in which the sender can tie their hands to punish if the target fails to make sufficient concessions. In this second model, the sender first chooses a level of concessions to demand, and ties their hands such that if the target fails to make these concessions, punishment will occur automatically. In this second model, coercion always works, assuming that the disputed good is continuously divisible.

These two models lead to my first conclusion. If the coercion scenario is isolated from other events, coercive success depends on the existence and form of commitment devices. If

the sender has some commitment device, such as tying their hands or audience costs, that would make it rational to carry through with the punishment threat, then coercion will be successful. However, if there is no such commitment device, then coercion will fail.

However, coercion scenarios will typically not happen in isolation. I thus reexamine both models of coercion in an infinite horizon context. When the actors time horizons are short, these models converge on the single stage models, and the existence and form of commitment devices remains the primary determinant of coercive success. However, when commitment devices are long, both models feature multiple coexisting equilibria. Equilibria where coercion succeeds coexist with equilibria where coercion fails. Commitment devices can no longer explain coercive success.

The existence of these multiple equilibria requires reference to constructivist factors to fully explain when coercion is successful. Determining when coercion is successful requires looking at intersubjective beliefs about whether coercion would be successful. In essence coercion would be successful when both sides believe it would be successful. These intersubjective beliefs would be socially constructed. In essence, when time horizons are long, coercion will work when and only when both sides believe it will work.

Thus, these models lead to two novel conclusions about coercion. First, time horizons are a critical factor in how we should explain coercion. Time horizons not only affect whether coercion will work, but change which factors will explain whether coercion will work. Second, coercive success may often be explained by socially constructed factors, rather than material interests. This second finding may have broader implications for international relations theory, as it shows that even the effect of hard power may be determined by socially constructed factors.

2 Background

For this paper, I define "coercion" as an attempt by one actor to get another actor to do or not do something by threatening to impose costs on the second actor. Importantly, I assume that the punishment threatened will do nothing directly to achieve the first actor's goals. Schelling (1966) noted that there was a distinction between using brute force to achieve a actor's objectives, and coercive strategies that manipulated the adversary's costs and benefits. Similarly, Snyder (1960) and Pape (1996) differentiated between strategies of denial, where one made it impossible for the adversary to achieve their objectives, and punishment, where one convinced the adversary to give in by threatening to impose costs. For this paper, I am looking only at punishment strategies.

Schelling (1966) also made a distinction between coercive strategies of deterrence and compellence. The former threatened punishment to convince an adversary not to do something, while the latter used threats to convince them to actively do something. George (1991) added an intermediate category of coercive diplomacy, where the threat was intended to get the adversary to reverse an action recently taken. However, all of these are part of coercion and follow the same basic strategic logic of using threats to convince an adversary to behave in a certain way. In addition, there are many cases where it is unclear whether a situation represents deterrence or compellence, such as U.S. attempts to coerce Syria to stop using chemical weapons (Moller 2013).

The dynamics of coercion have been studied both informally and formally for decades. While some papers have studied coercion in general (e.g. Schelling 1966; Slantchev 2003; Jervis 1979; Lebow and Stein 1989), others have focused on specific means of coercion, such as nuclear war (Powell 1990; Powell 2015; Anderson, Debs, and Monteiro 2019; Kronig 2013; Beardsley and Asal 2009; Sechser and Fuhrmann 2013), coercive bombing (Pape 1996), sanctions (Tsebelis 1990; Smith 1996; Drezner 1998; Lacy and Niou 2004; Krustev 2010; Whang, McLean, and Kuberski 2013; Miller 2014), and terrorism (Abrahms 2011; Trager and Zagorcheva 2006). Thus coercion applies to a number of common scenarios in inter-

national relations. While papers examining coercion informally or observationally provide valuable insights, formal models are necessary to ensure that theories of coercion are logically sound. In addition, observational studies are limited by selection bias, as both successful and unsuccessful coercion may go unobserved (Danilovic 2001; Fearon 2002).

However, most previous formal models have important limitations. One issue common across almost all models of coercion is that they model only a single round of coercion, ending when one side gives in.[1] However, international relations is nearly universally characterized by repeated play. Even if the exact same situation does not repeat, there will be enough similar situations in the future that states and other actors must act as if they were in a repeated game. Substantively, the failure to model repeated play eliminates the ability to model or study the building or effect of reputation on coercion. This is critical, as reputation is one of the key factors that could make coercive threats credible.

One commonly used model is the chicken game (e.g. Schelling 1966; Zagare 1990; Kilgore and Zagare 1991), where each player simultaneously makes the choice to give in or stand firm. This game has three important limitations. First, actions are chosen simultaneously, whereas coercion dynamics inherently require the threat sender to react to the target's behavior. Second, it is unclear how the actions in the coercion game map onto substantive choices. For instance, the sender's choice to stand firm would mean inflicting punishment if the target fails to comply, but something different if the target does comply. Third, since each player has the same choices, the chicken game assumes the game is symmetrical. Instead, coercion dynamics have a clear threat sender and threat target, unless one is trying to model mutual coercion.

Other models do attempt to build a more realistic account of coercion. However, they still tend to suffer from several limitations in addition to limiting coercion to a single round. In particular, most models tend to focus on a specific situation or type of coercion, and so it

[1]Some models do include repeated moves, but still end when one side gives in and/or when punishment (e.g. a nuclear attack) occurs (e.g. Powell 1990). They thus still model only a single coercive situation, and have the same substantive issue as models with only a single round of play.

is unclear how generalizable they are. How well they model that particular situation can also be debatable. For instance, models of nuclear coercion (Powell 1990; Powell 2015) commonly assume a brinksmanship format, such that there is some possibility a nuclear attack could occur by accident. However, it is unclear whether accidental punishment is realistic even in the nuclear realm, let alone extending to other types of coercion.

A second common issue is that many models (Tsebelis 1990; Drezner 1998; Lacy and Niou 2004) simply assume that the target will carry through with their punishment threat, at least some of the time. In these models, the sender prefers to punish if the target fails to make appropriate concessions. However, punishment threats would generally be at least somewhat materially costly to the sender as well. Thus, this assumption is problematic. In addition, these models also assume that the sender has no punishment option if the target does make concessions. If the sender benefits from punishing, they also need to be able to credibly commit not to punish if the target makes concessions, rather than this be assumed. Thus, models need to both allow a punishment option even if the target makes concessions, and critically examine the sender's preferences over punishment.

A third limitation is that most models do not include the potential for a continuous offer, thus assuming that the disputed good or object is indivisible, although the substantive implications of this assumption are unclear.

Finally, some models (Signorino and Tarar 2006; Tragar 2013; Kydd and McManus 2017; Slantchev 2005; Sechser 2010) that claim to model coercion are not actually modeling coercion as I have defined it. These models assume that the sender has some possibility of getting what they want through direct action, even without the target's compliance. There are situations in which this is an appropriate assumption, such as conventional war, and it is potentially valid to label these situations as coercion. However, by making this assumption these models are studying something fundamentally different than the types of coercion by punishment that I am discussing.

3 Coercion models in a finite setting

To understand how various factors affect coercive success, I will first examine models that have a single or otherwise finite number of stages. These models would represent cases where the coercive scenario occurs rarely and there is little connection to other coercion scenarios. In this case, the primary factor determining whether coercion can be successful is the ability of the two sides to either tie their hands or otherwise lock-in an incentive to execute a certain choice, even if it would no longer be in their interests. In particular, if the threat sender can tie their hands to punish, even if they receive no concessions, coercion will generally work. In contrast, if the sender cannot tie their hands or the target can tie their hands first, coercion will fail. Similar results would occur if the sender has some ability to incentivize carrying out the punishment threat, even if it is costly and will not generate concessions. For instance, perhaps the sender could generate audience costs.

Below I will examine two basic models of coercion. The first is a basic coercion model, with no additional ability of either side to tie their hands. In the second model, I allow the sender to tie their hands such that punishment occurs automatically if a given amount of concessions are not made. Coercion always fails in the first model, and always works in the second model. This demonstrates how the model form, and hence which strategies are available to the players, determines whether coercion works in finite games. In the online appendix, I will consider other models, such as one where the sender can generate audience costs. These models are essentially elaborations of these two models, and confirm that explaining coercive success in finite games depends on explaining which strategies are available to the actors.

3.1 The basic coercion model

I begin by examining a basic model of coercion, with no additional strategies to tie-hands or incur audience costs. This model serves as the basic foundation for all other models.

In the model, the sender issues a coercive threat that cannot directly achieve any of the sender's goals. This basic model form is thus synonymous with coercion by punishment. If a situation is not well represented by this model or a variation on this model, it is not coercion by punishment.

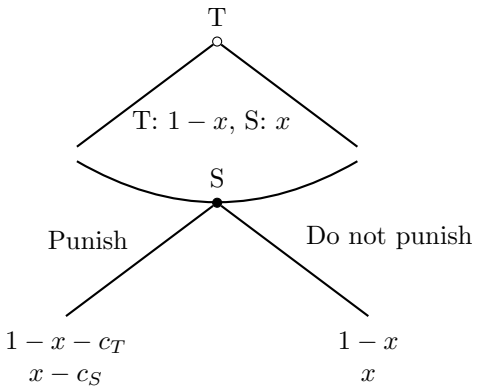
In this model, I assume that there are two actors, a sender (S) of a coercive threat and target (T) of the threat, in a dispute over some continuously divisible good or policy. This good has a total normalized value of 1, and is completely rivalrous, so each actor only gets benefits from the portion of the good that they possess. The target will decide how to divide the good in each period, giving the sender x concessions, and retaining $1 - x$. The central feature of pure coercion scenarios is that carrying out a coercion threat cannot capture any of the good directly. Therefore, the division of the good proposed by the target is implemented regardless of the sender's actions.

While the sender cannot directly affect the division of the good, I do allow the sender to carry out their coercive threat and punish the target. Specific punishment actions could include economic sanctions, coercive bombing, or terrorist attacks among other actions. If the sender chooses to punish, the punishment imposes costs on both the target and sender, c_T and c_S respectively.[2] Note that these cost parameters actually represent the combination of several elements. First, they include the actual costs that punishment would inflict. Second, where costs may be probabilistically inflicted, the cost parameter represents the overall expected costs, and thus includes the probabilities of different amounts of costs (including no costs) being inflicted. Finally, the parameter incorporates the amount that each side cares about these costs relative to the value of the disputed good, commonly labeled resolve. Thus, if the sender chooses to punish, the target's utility is $x - c_T$, while the sender's is $1 - x - c_S$. If the sender chooses not to punish, the division of the good is implemented

[2]I do assume that carrying out punishment inflicts at least some costs on the sender. I believe this is appropriate, as the sender would have to pay some costs (even if minimal) to carry out any conceivable punishment action. In addition, the costs to the sender could include natural retaliation, such as counter-sanctions or retaliatory bombardment. In the online appendix, I will consider a scenario where the sender directly benefits from punishing the target.

without modification. The basic structure of the game is shown in figure 1.

Figure 1: The Basic Coercion Game



Note that this model leaves the communication of the coercive threat and all subsequent communications implicit. This means that the model assumes that communications are cheap talk and do not have any direct effect on the actors utility functions. Overall, this seems like a reasonable assumption, although I do include a model with an explicit threat that creates audience costs in the online appendix.[3]

Since this is a finite, perfect information game, I adopt the solution concept of subgame perfect Nash equilibrium (SPNE). Thus the model can be solved through simple backwards induction. As long as there are a finite and specific number of stages, the solution will be the same as if there were only one stage. The one stage solution would be the solution to the final stage, which would prevent any punishment strategies. Thus, all previous stages would also have the same solution as the final stage. In addition, if there were more than a few stages, the situation would be better modeled by an infinite horizon game, as I will model later in the paper. Therefore, I will discuss the model solution as if it had only a single stage.

In the basic coercion model, the only equilibrium is one where coercion never works. To be able to compare this across models, I will label any equilibrium where coercion fails a no-concessions equilibrium.

[3]Not available yet

If the sender experiences positive costs from coercion, the sender will never punish. In finite versions of the model, the sender makes the final move. By definition, punishment does not affect the distribution of the disputed good and does nothing but impose costs on both parties. Thus, punishment only lowers the senders utility, and the sender will never punish regardless of whether the target made concessions. In the basic coercion model, the punishment threat is inherently incredible. Knowing that the sender will never punish, the target will never make any concessions in the first place. Lemma 1 displays the equilibrium of the basic coercion model when the sender experiences positive costs.

Lemma 1. *The no-concessions equilibrium in the finite basic coercion model*

In the basic coercion model, the target will offer $x = 0$, making no concessions, and the sender will never punish.

This is the only equilibrium in finite versions of the basic coercion model

Thus, in the basic coercion model, coercion is never successful.

The basic coercion model made two assumptions about the form of the game. However, neither is material to the equilibrium. I assumed that the disputed good was continuously divisible, and I assumed that the sender had a limited choice to punish or not punish rather than choosing an amount of punishment. Even if one or both of these assumptions were violated, the basic equilibria will hold and coercion will never work. Even if the sender can choose the amount of punishment, punishing at all will either inflict costs on the sender. The sender will still have to go last, and punishment will not otherwise affect the outcome. In the first case, the sender will never be able to credibly commit to inflict any punishment. Thus, the target's optimal move will always be to make no concessions. Since the target will never make any concessions, it does not matter whether or not the disputed good is divisible.

3.2 The coercion model where sender can tie hands

However, it is possible that the sender has some way to tie their hands or otherwise incentivize themselves to behave in a way they otherwise wouldn't. To examine this possibility, I look at

an extreme version of the sender's ability to tie hands. In this model, the sender commits for certain to punish if a certain amount of concessions are not made. At that point, punishment is carried out automatically. Other models that include some ability to tie hands or create incentives will be a variation on either or both of the basic coercion model or the tying hands model. For instance, the brinksmanship models common in studying nuclear deterrence are simply a milder version of the tying hands model. In those models, instead of making punishment happen automatically and for certain if concessions are not made, the sender ties there hand to some random chance of punishment happening automatically.

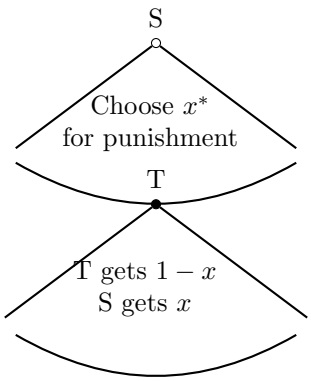
In the tying hands model, I again assume that there are two actors: the sender (S) of the coercive threat and the target (T) of the threat. Again, they are in a dispute over some continuously divisible good with a total normalized value of 1. Again, the target retains complete control over the distribution of the disputed good, giving the sender x concessions and retaining $1 - x$ portion. I do assume that the disputed good is continuous, and that the target can divide it however they wish. Unlike in the basic coercion model, indivisible disputed issues may affect the outcome of the game.

However, unlike the previous model, the sender moves first by deciding whether and how to tie their hands. The sender makes a choice of the amount of concessions x^* the target must make to avoid punishment. Thus, punishment will automatically occur if the target makes insufficient concessions, offering $x < x^*$. Punishment will never occur if the target does make sufficient concessions, offering $x \geq x^*$. Since the sender has tied their hands as to whether punishment will occur, the sender makes no further decisions. I assume that the target observes the sender's tying hands behavior, and knows the amount of minimum concessions they need to make in order to avoid the automatic punishment (i.e. the target knows the x^* the sender has chosen.

Once the sender has made their decision on how to tie their hands on whether to punish, the target decides the allocation of the good. At that point punishment is either inflicted on both parties, or not depending on whether the target has made sufficient con-

cessions. As above, if punishment is inflicted, both sides suffer the costs of punishment, c_T and c_S. Figure 2 shows the game tree for the model where the sender ties their hands about whether punishment will occur.

Figure 2: Coercion Game Where Sender Ties Hands



If $x \geq x^*$: T receives $1 - x$, S receives x
If $x < x^*$: T receives $1 - x - c_T$, S receives $x - c_S$

Since this is again a finite, perfect information game, I again adopt the solution concept of subgame-perfect Nash equilibrium. Again, I will focus on the situation where there is only a single stage, although the solution will hold for any specific number of stages.

In contrast to the basic coercion model, in the tying hands model, coercion always works at achieving some concessions, as long as the disputed good is continuously divisible and punishment is mutually costly. I will label any equilibrium where coercion is successful a concessions equilibrium.

In the tying hands model, the target goes last. I also have assumed that they observed at what level x^* the target has chosen for punishment to automatically occur. Thus, the target will make concessions if the concessions needed to avoid punishment are less than their cost of being punished $x^* \leq c_T$. If the concessions needed are more than the cost of being punished, then the target will make no concessions, preferring to accept the costs of punishment.

The sender's knowledge of this shapes their behavior. The sender will tie their hands such that coercion is always successful. The sender will choose the maximum level of con-

cessions that the target is willing to make before the target would accept punishment. Thus, the sender will choose $x^* = c_T$, automatically inflicting punishment if the target makes fewer concessions than the target's costs of being punished. The sender will not tie their hands to a higher level of concessions, as the target would accept punishment rather than make concessions, thus inflicting costs on both parties. The sender has no reason to tie their hands to a lower level of concessions, as this would just reduce their benefits. Note that the sender's costs of inflicting punishment do not matter, as the sender knows that the target will make the needed concessions, and so the sender does not have to worry about punishment actually being implemented. Lemma 2 displays the concessions equilibrium for the finite tying hands model.

Lemma 2. *The concessions equilibrium in the finite tying hands coercion model*

In the tying hands model, the sender will always choose to automatically punish at the concessions level $x^* = c_T$, within punishment occuring if the target offers $x < x^*$ and no punishment occurring if the target offers $x \geq x^*$

The target will always make concessions, offering $x = x^* = c_T$

This is the only equilibrium in finite versions of the tying hands model.

Thus, in the tying hands model, coercion is generally successful.

Unlike the basic coercion model, an indivisible disputed goods may affect the solution somewhat. Recall that the target will only make concessions up to the costs they would pay in being punished. If the disputed good is indivisible such that the minimum concessions the target could make outweigh their costs of being punished, the target would prefer to make no concessions and accept the costs of punishment. If punishment is mutually costly, the sender would thus not tie their hands, and punishment would not occur. Thus, in the tying hands model whether the disputed good is continuously divisible can affect whether coercion is successful.

3.3 Discussion

Two major findings emerge from the finite versions of the coercion models. First, these models show that all punishment threats have a credibility problem, where it would be irrational to carry out the punishment threat absent a commitment device in finite games. The problem of punishment threats being incredible is not limited to nuclear threats. Second, the models show that the exact form of the model is crucial to explaining coercive success in finite settings. In particular, whether coercion is successful is heavily dependent on whether the actors have some commitment device to tie their hands or otherwise incentivize themselves to carry out an action that would otherwise be irrational.

A leading issue in early studies of nuclear was how to make a nuclear threat credible, when carrying out that threat would lead to mutual annihilation, and thus be irrational. The common chicken analogy and brinksmanship models were thus premised on carrying out the punishment threat being mutually disastrous, and not merely mutually costly. However, the basic coercion model shows that the issue of punishment threats is universal as long as punishment is mutually costly. Even if the target would incur far more costs than the sender (relative to the value of the disputed good), carrying out the punishment threat would be irrational, and thus the coercive threat is incredible.

The finding that carrying out a punishment threat inherently presents credibility concerns is likely not entirely novel. As mentioned, it was inherent in studies of nuclear coercion. Most studies on other forms of coercion by punishment have likely encountered the difficulty of making a credible coercive threat when carrying out the threat is mutually costly, and thus irrational.

However, it is still useful to reiterate that all punishment threats face this issue. It would either be irrational to carry out the threat or irrational to refrain from carrying out the threat. This is true in all finite games that lack some incentive structure to make both carrying out the threat if no concessions are made.Reiterating this point provides the foundation for the rest of the analysis in this paper.

A second general finding worth reiterating is that the target will never make concessions greater than their cost of being punished. If the value of the concessions demanded outweighs the cost of punishment, the target would prefer to accept punishment rather than make ay concessions. If the disputed issue is continuous, this means the maximum concessions the sender can get equal the target's cost of being punished. If the disputed issue is indivisible, the sender can only gain concessions if the target values the disputed issue less than the cost of being punished.

Third, these issues mean that explaining coercive success in finite games requires determining if the sender has some sort of commitment device that will either tie their hands to punish in certain circumstances, or provide some inducement that would lead them to punish when it would otherwise be irrational. As noted above, punishment threats face an inherent credibility issue, where it would be irrational for the sender to carry out their punishment threat as long as punishment is mutually costly. In order to gain concessions, the sender needs some sort of commitment device to overcome this problem.

One form of commitment device, illustrated in the model above, is for the sender to tie their hands to punish or not punish before they know whether the target will make concessions. Other forms are illustrated in the online appendix.[4] For instance, a brinksmanship strategy would be a softer form of tying hands. In essence, brinksmanship simply involves the sender tying their hands to punishing with some random probability rather than for certain. Another commitment device might involve making a public threat such that the sender would incur audience costs if they fail to punish when the target does not make concessions or go ahead and punish when the target does make concessions. In this case, the audience costs might override the costs of punishment, providing an inducement to go ahead and punish even though punishment provides no other benefits.

However, whether a commitment device exists to make the punishment threat credible depends on the exact model examined. There is no commitment device in the basic coercion

[4]Not available yet

model. Thus, the existence of a commitment device depends on which model is the best fit for a given situation and which strategies that situation allows. The existence of commitment devices, and thus whether coercion is successful will be highly situation specific.

An example that might make this situation specificity clear is the Persian Gulf War. In this case, Iraq was attempting to use coercion to maintain control of Kuwait. Iraq could have no hope of fully defeating the coalition forces, and defeat was certain. This thus represents a coercion through scenario where Iraq was trying to deter a coalition offensive by threatening to impose costs on the coalition. In this case, Iraq did have a commitment device where they could essentially tie their hands. By deploying their forces on the border, any coalition forces would have to fight the Iraqi forces, and thus incur the costs of fighting. Even though victory was certain, the coalition would be punished in their victory. In this case, the threatened punishment was insufficient to deter the coalition forces, but Iraq did successfully tie their own hands to fight.[5]

However, Iraq would not have been able to tie their hands to punish in a coercion attempt to regain Kuwait after the war. In this case, Iraq would have had to actively order their forces to attack, knowing they would be defeated. While such an attack would impose costs on the coalition, it would also impose costs on Iraq. Since Iraq would have to actively order their forces forward, they would be unable to tie their hands as they successfully did in trying to defend Kuwait. Accordingly, any Iraqi punishment threat after the war would not be credible.

The impact of commitment devices will be even more situation specific as the target may have their own commitment devices. For instance, the target may be able to commit to a certain level of concessions (including none). If both sides have a commitment device, whether coercion would be successful would depend on which side could commit first. If the target could tie their hands first, the situation would essentially become the basic coercion

[5]Uncertainty also plays a role, as Iraq did believe that punishment would inflict sufficient costs on the coalition to deter the offensive. The actual punishment results from this miscalculation. However, the miscalculation would have been irrelevant if Iraq did not have the commitment device of deploying their forces.

model. If the sender tied their hands first, the situation would remain the tying hands model illustrated above. Other commitment devices might have a more complicated interaction. However, which model was most accurate, and thus whether coercion would be successful, would still be highly situation specific.

It is worth emphasizing that the primary explanatory factor is the form of the model and what strategies are available to the actors, and not the parameters in the model. In some cases, such as the two models presented above, parameters are largely irrelevant to whether coercion is successful. In other models, such as the audience costs model in the appendix, parameters can affect coercive success. However, the parameters only matter after a specific model form has been determined. Empirically examining coercive success must place priority on the strategies available to the players over parameters in the model, such as the costs of punishment. Simply plugging common explanatory variables, such as power or regime type, into a regression model is unlikely to be successful at explaining coercion.

4 Coercion models in an infinite horizon setting

The previous discussion assumed that each coercion scenario happened in isolation, or at least in connection with only a small number of similar scenarios. In other words, both actors in the coercion scenario were concerned only about the immediate events. In each of the models, the coercion scenario occurred only a finite and definite amount. Really, this only makes sense if each coercion scenario was completely unique, and the actors were not concerned about how their actions in the present coercion scenario would affect other events.

However, I believe that most of the time coercion is better modeled as a repeated game that actors play an indefinite number of times, especially in international relations. Rarely do actors disappear from the world stage, and thus act repeatedly with the same actors. In addition, they will often encounter similar situations in dealing with other actors. While the exact circumstances of a coercion scenario may not repeat, at least somewhat similar

situations are likely to occur. Thus, in making their decisions within a given interaction, the actors must consider not only the immediate situation but how their actions will affect future interactions. This leads to modeling coercion as a repeated game with an indefinite or infinite number of stages.[6]

Infinite horizon models also make it possible to study the effects of reputation on coercive success. One reason the sender might punish, even when punishment provides no immediate benefits and is costly, is to maintain a reputation for fulfilling their punishment threats. This could allow the sender to gain concessions in the future, outweighing the current costs of punishment. On the other hand, the target may also want to maintain a reputation for intransigence in order to avoid future coercive threats.

I thus reanalyze both of the basic models presented above in an infinite horizon setting. Thus, each of the models is played repeatedly an indefinite number of times. The infinite horizon versions of the model vastly change the conclusions of the finite versions of the model, and lead to some novel insights about coercion. In the infinite horizon versions of the model, the actual model form generally does not play a significant role in whether coercion works. Most models feature multiple sensible equilibria, some in which coercion is successful and some in which coercion fails. Thus, the existence of commitment devices no longer plays a significant role in determining coercive success. The sender's concerns about reputation often are sufficient to make a costly punishment threat credible. At the same time, the target's reputation concerns can override the impact of the commitment device, leading the target to refuse concessions even knowing they will be punished.

The existence of multiple equilibria means that coercive success is determined by which equilibrium is actually played. I argue that this equilibrium selection requires reference to socially constructed beliefs and norms. Assuming a sufficient shadow of the future, coercion will work when and only when both sides believe it will work.

[6]Mathematically a finite but random number of stages is modeled the same as a game with infinite stages. In the former case, the discount parameters simply also capture the probability of the game ending.

4.1 Basic coercion model - infinite horizon

I start by analyzing an infinitely repeated version of the basic coercion model described above. Recall that in this model, the target first divides the disputed good, offering the sender x portion and retaining $1 - x$. The sender then has the option to punish, in which case both sides suffer punishment costs c_T and c_S. The infinite horizon model is identical, except this basic framework is repeated an infinite number of times. Each side discounts future stages at the rate of δ_T and δ_S.

Since this is still a perfect information extended model, I continue to use the solution concept of subgame perfect Nash equilibrium. While the folk theorem states that there will be an infinite number equilibria, these can be divided into two broad equilibrium classes. There is always a no-concessions equilibrium, where the target makes no concessions and the sender never punishes. For some set of parameter values, there is also a set of equilibria where the target does make concessions under the threat of punishment, and the sender refrains from punishing as long as appropriate concessions are made. Any other equilibrium will have to include elements of one or both of these equilibria.

The no-concessions equilibrium is identical to no-concessions equilibrium in the finite versions of the basic coercion model, where it was the only equilibrium. This equilibrium is stated in lemma 3. The logic is also the same as in the finite model. If the sender does not expect punishing to induce concessions in the future, punishment simply lowers the sender's utility. Therefore the sender would have no reason to punish. Knowing this, the target can safely refuse to make concessions, retaining the entire disputed good. The target's strategy of refusing to make concessions then confirms that punishment will not induce future concessions, thus confirming the sender's strategy of never punishing. In essence, the carrying out the punishment threat may still be irrational in the infinite horizon version, which would make it irrational for the target to make concessions.

Lemma 3. *: The no-concessions equilibrium when in the infinitely repeateed basic coercion model*

In the infinitely repeated basic coercion model, there is always an equilibrium where the target offers $x = 0$, making no concessions as follows, and the sender never punishes.

Note that as in the finite model, the no-concessions equilibrium exists regardless of the relative costs of punishment. The sender's incentive to punish is eliminated solely because S incurs costs by punishing. Because S suffers costs in punishing, S has no incentive to actually punish unless doing so would lead to future benefits. Because T is not offering any future benefits in response to S's punishment, S will not attack. Thus, the no-concessions equilibrium exists even if punishment inflicts far greater costs on the target than on the sender.

In the finite version of the basic model, the no-concessions equilibrium was the only equilibrium. In infinite horizon versions, it is possible that there is a range of equilibria where coercion succeeds. The basic concessions equilibria are detailed in lemma 4.

Lemma 4. *: The concessions equilibrium in the infinitely repeated basic coercion model*

In infinite horizon versions of the basic coercion model, any division of the disputed object $c_T \geq x^* \geq \frac{(1-\delta_S)c_S}{\delta_S}$ can be sustained with the following strategies:

T always offers x^*, unless S has failed to punish when T makes an offer $x < x^*$, in which case T offers $x = 0$ from then on.

S punishes if T makes an offer of $x < x^*$, and does not punish if T makes an offer of $x \geq x^*$

In the concessions equilibria, the sender is willing to punish. Therefore, the target is willing to make concessions as stated in the specific concessions equilibrium, as long as those concessions are less than the target's costs of being punished. Punishment is still costly to the sender in the short term. However, if the sender fails to punish, both actors would divert to the no-concessions equilibrium. The sender's punishment threat is thus made credible by the need to preserve future concessions.

In essence, these equilibria can be interpreted as the sender having a reputation for punishing if their demands are not met. This reputation is enough to induce the target to

make concessions. The sender is willing to punish in order to preserve that reputation, which makes the sender's punishment threat credible.

However, note that the the no-concessions equilibrium still exists. Thus, if any concessions equilibria exist, they coexist with the no-concessions equilibrium. Accordingly, explaining coercive success in infinite horizon versions of the basic coercion model requires explaining which of these equilibria is actually played and why. I will return to these points after examining the tying hands model.

4.2 Tying hands model - infinite horizon

Next, I examine the tying hands model, which represents the opposite extreme with the sender having a fully effective commitment device. Recall that in this model, the sender ties their hands to punish at some level of concessions x^*. If the sender makes any fewer concessions (offers $x < x^*$), punishment happens automatically. If the sender makes sufficient concessions (offers $x \geq x^*$), punishment will never occur. After observing the sender tying their hands, the target then chose the level of concessions they would actually make. At that point punishment would occur or not according to how the sender tied their hands relative to the level of concessions. The infinite horizon version is simply the above game repeated infinitely. As above, future rounds are discounted by δ_T and δ_S.

Since the sender chooses how to tie their hands each round, I do assume that the sender can only tie their hands temporarily. Either the tying hands mechanism is inherently temporary, or can be undone with sufficient fore-notice. I believe this is reasonable, as it is difficult to think of a mechanism in which the sender could tie their hands permanently and irrevocably.

I again adopt the solution concept of subgame-perfect Nash equilibria, as this is again a perfect information extended game. The same two basic equilibria classes exist in this model as in the infinite horizon basic coercion model. The only difference is that in this case it is the concessions equilibrium that always exists, while the no-concessions equilibrium

only exists if the target has a sufficiently long time horizon.

The concessions equilibrium in the infinite horizon tying hands version is identical to the concessions equilibrium in the finite version of the model. This equilibrium is stated in lemma 6. The logic also follows that of the finite version of the model. The target knows that the sender has tied their hands to punish if the target does not offer the demanded concessions. It is thus in the target's interest to make the demanded concessions, as long as they are less than the costs the target would receive if they were punished. Knowing this, the sender will demand the maximum amount of concessions the target is willing to make, and tie their hands to automatically punish if these concessions are not made.

Lemma 5. *The concessions equilibrium in the infinitely repeated tying hands coercion model*

In the infinitely repeated tying hands model, there is always an equilibrium where the target makes concessions $x = c_T$ as follows: the sender will always choose to automatically punish at the concessions level $x^* = c_T$, with punishment occurring if the target offers $x < x^*$ and no punishment occuring if the target offers $x \geq x^*$. The target will always make concessions, offering $x = x^* = c_T$

While this concessions equilibrium was the only equilibrium in finite versions of the model, there can be a no concessions equilibrium in infinite horizon versions. This no-concessions equilibrium is detailed in lemma 6.

Lemma 6. *The no-concessions equilibrium in the infinitely repeated tying hands coercion model*

In the infinitely repeated tying hands model, there is an equilibrium where the target makes no-concessions as long as $\delta_T \geq \frac{1}{2}$. This equilibrium uses the following strategies:

S commits not to punish ($x = 0$) in all rounds. T makes no concessions, regardless of the concessions S demanded.

If T ever makes concessions, S will demand $x^* = c_T$ or $x^* = 1$, whichever is less, and

commit to punishing if T offers $x < x^*$ and not punishing if T offers $x \geq x^*$. T then offers $x = x^*$ in all subsequent stages.

In this equilibrium, the target never offers concessions on-path, and the sender always ties their hands to no punishment. Thus, the target can safely make no concessions as they will not be punished. The sender knows that the target will still not make concessions if they demand some level of concessions and tie their hands accordingly. This will simply result in both sides being punished, reducing the sender's utility. Accordingly, the sender will never tie their hands to punish.

However, for both this equilibrium to hold, the target must be able to credibly commit to resist making concessions even if the sender demands concessions and ties their hands to punish. The target can resist making these concessions because they know making concessions will cost them more in the long run, even if it avoids punishment in the short term. If the target does make concessions, both sides divert to the concessions equilibrium, with the sender always demanding concessions and tying their hands and the target complying. To avoid making many concessions in the future, the target resists making concessions in the present, even if they would be punished.

This equilibrium can be interpreted as the target refusing to make concessions in order to retain a reputation for steadfastness and refusing to back down. Maintaining this reputation means that the target can avoid concessions in the future. However, if the target ever backs down and makes even minor concessions, their reputation will be ruined, and they will have to make many concessions in the future. Thus, the target is willing to suffer punishment in the short term to maintain this reputation.

However, note that the concessions equilibrium always exists. While the no-concessions equilibrium may exist for a fairly wide range of parameter values, it would coexist with the concessions equilibrium.[7] Thus, the infinite horizon version of the tying hands model also

[7]There may also be additional concessions equilibria, where the target gains fewer concessions than the maximum that the target is willing to make (i.e. $x^* < c_T$). These would have similar strategies to the no-concessions equilibrium.

often has multiple equilibria. Explaining whether coercion works or fails would require explaining which equilibrium is actually chosen.

4.3 Multiple equilibria and coercive success

A few major findings emerge from the infinite-horizon models. When actors have long enough time-horizons, whether there is a clear commitment device no longer matters to coercive success. Because there are multiple equilibria, explaining whether coercion is successful requires explaining which equilibrium will occur.

Before elaborating on these points, recall that coercive success is synonymous with the actors playing a concessions equilibrium. In the concessions equilibria, the target makes some concessions. Thus, coercion was successful. Similarly, coercive failure is synonymous with the actors playing the the no-concessions equilibrium. In that case, the target never makes any concessions, and therefore the coercive threat has failed.

Each finite game had only one equilibrium. Thus, coercive success or failure was dependent on determining which game is being played. This in turn depended on whether the sender had some sort of commitment device to make there punishment threat credible, despite punishment being costly and providing no immediate benefits. The exact form of the commitment device and the parameter values may further determine whether coercion will work. When the actors have relatively short time horizons, the infinite horizon games also have only one equilibrium. With short time horizons, coercive success is still dependent on the form of the game, and thus still dependent on whether a commitment device exists and the form of that commitment device.

However, with sufficiently long time horizons, the models have multiple equilibria, some in which coercion works and some in which coercion fails. Note that the existence of these multiple equilibria occurs for the exact same parameter values. In addition, the existence of multiple equilibria extend to other models of coercion, such as the brinksmanship

model or models with audience costs, provided the actors have long enough time horizons.[8] Thus, when coercion scenarios are repeated and with sufficiently long time horizons, coercive success is no longer explained by the existence of commitment devices. In fact, it is possible that commitment devices would no longer play any role in determining whether coercion works.

Why don't commitment devices play a significant role in determining coercive success when games are repeated with sufficient time horizons? The answer is that the actors concerns about reputation and the associated future benefits outweigh the impact of short-term commitment devices. Provided it will support a reputation giving them future concessions, the sender is willing to punish even without any commitment device. Similarly, the target may refuse to make concessions even knowing punishment will occur in order to maintain a reputation that will allow them to avoid concessions in the future. Thus, with long enough time horizons, commitment devices are no longer necessary for the sender to gain concessions and no longer sufficient to force the target to make concessions.

Accordingly, in repeated games with long enough time horizons, coercive success is explained completely be determining which of multiple equilibria actually occurs. Given that there are multiple equilibria, the exact parameters do not provide significant guidance on which equilibrium will actually occur, provided the actor' time horizons are sufficiently long. Given that these multiple equilibria occur across a wide variety of models, the exact strategies available to the actors also do little to determine whether coercion will work. Thus, it appears that neither the specific form of the game nor the parameters of the model are sufficient to explain coercive success with sufficient time horizons.

[8]See appendix X (doesn't exist yet).

4.4 Socially constructed beliefs, equilibrium selection, and coercive success

The previous section showed that in repeated games with sufficiently long time horizons, multiple sensible equilibria coexisted. Whether coercion is successful depends entirely on which of these equilibria is actually played. Thus, explaining coercive success means explaining equilibrium selection. Political scientists and formal modelers do not yet have a convincing explanation for why one equilibrium may be more likely to occur when a game has multiple equilibria.

To begin to examine which equilibrium is more likely, and how constructivist beliefs can explain equilibrium selection, I believe that it is necessary to return to the original definition of a Nash equilibrium. John Nash defined an equilibrium as a set of strategies "such that each player's mixed strategy maximizes his payoffs if the others are held fixed. Thus each player's strategy is optimal against those of the others (Nash 1951)." In other words, each player's strategy is a best-response to all the other player's strategies. For an equilibrium to occur, each player must correctly anticipate what strategy the other will play, and choose their best response. The other player's strategy is in turn determined by correctly anticipating the first player's strategy and also choosing their best response.

Therefore, each equilibrium requires that the players share beliefs about which equilibrium will occur and what strategies they and the other players will play. In turn, this means that the actual equilibrium played will be determined by these intersubjective beliefs. If the players share a belief that one equilibrium will occur, then that equilibrium is the one that will actually occur. If the player's intersubjective beliefs dictate that a different equilibrium will occur, then that second equilibrium is in fact the one that will occur.

These intersubjective beliefs must be socially constructed. Because there are multiple equilibria, there is no material basis to support the actors playing one equilibria rather than another. The only way to form these intersubjective beliefs is by thinking about socially constructed factors and what each actor expects to happen.

To apply this to the coercion model, the no-concessions equilibrium will be played if the two actors believe that it will be played. Similarly, a concessions equilibria will be played if the two actors believe that a concessions equilibria will be played. In the later case, the exact equilibrium within the possible range, i.e. the exact amount of concessions will be similarly determined by the players' intersubjective beliefs. In essence coercive threats will be successful if both sides believe they will be successful, and fail if both sides believe they will fail.

Thinking through the logic of coercion informally also reveals the importance of intersubjective beliefs in determining coercive success in repeated games.

First note two straightforward statements that are clearly true. First, it would only be rational for the target of a coercive threat to make concessions if they believes they will be punished for not making concessions. Second, it is only rational for the sender to actually punish (assuming punishment is mutually costly) if the sender believes that punishing will induce the target to make concessions in the future. These two statements show that each side's actions depend on their beliefs about what the other will do.

Given these statements, each side can infer the conditions under which the other will act. The target knows that sender will only punish if the sender believes that punishing will induce the future target to make concessions. Therefore, it would only be rational for the target to make concessions now if the target believes that the sender believes that punishing will induce the future target to make concessions. At the same time, the sender knows that the future target will only make concessions if the future target believes that the future sender's threat is credible. Therefore, it is only rational for the sender to punish if the sender believes that carrying out the punishment threat will make the future target believe that the future sender's threat is credible. Here, we can see that not only do both sides have to have beliefs about the future actions of the other, they need to have beliefs about the beliefs of the other.

The actors can also trace the chain of beliefs further. The target knows that the

sender will only punish if the sender believes that the next stage target will believe that the sender's future threats are credible. Therefore, the target will only make concessions if the target believes that the sender believes that the (next stage) target believes that the (next stage) sender's coercive threat will be credible. Similaly, the sender will only punish if they believe that the future target will therefore believe that the future sender's threat is credible. This means that the sender will only have a credible coercive threat if the sender believes that the (next stage) target will believe that the (next stage) sender believes that the (third stage) target would give in to the coercive threat.

Each actor now needs beliefs about the other's beliefs about their own beliefs. This process can be repeated indefinitely, with the credibility of the coercive threat depending on beliefs about beliefs about beliefs about beliefs and so on. In essence, it is now clear that coercive threats depend on intersubjective beliefs. As previously noted, if both sides believe the coercive threat will work, then it will actually work. If both sides believe the coercive threat will fail, then it will actually fail.

5 Explaining coercive success: putting it together

The previous findings provide a general guide for how to explain when coercion will be successful. While this is not necessarily exhaustive, it should provide a framework for future theoretical and empirical analysis.

To predict whether coercion is successful in a given circumstance, researchers should first identify how frequently the game is repeated and the time horizons of the relevant actors. If the game is not repeated and / or the actors have short time horizons,[9] then the existence and form of commitment devices will be the primary determinant of whether coercion will work. The researcher could then proceed to examine the specific game form and existence of commitment devices to predict whether coercion will work. If the game is

[9]In some sense infrequent repetition and short time horizons are synonymous. If the coercion scenario occurs infrequently, by the time it will reoccur so much time will pass that the actors can be presumed to heavily discount the gains from the next interaction in the far future.

frequently repeated and the actors have long time horizons, then the primary determinant of whether coercion will work will be socially constructed intersubjective beliefs about whether coercion will work. The researcher can then focus on determining what will influence these beliefs.

If one actor has a long time horizon and one actor has a short time horizon, whether commitment devices or intersubjective beliefs explain coercion will depend on the interaction between which actor has the long time horizon and the existence of commitment devices. If the sender has the long time horizon, there would be multiple equilibria when there are no commitment devices, and hence intersubjective beliefs would better explain coercion. However, in cases where the sender has a long time horizon, the target has a short time horizon, and the sender can use commitment devices to ensure the credibility of the punishment threat, the only equilibria would be the concessions equilibrium, and the commitment device would best explain coercive success. Conversely, if the target has a long time horizon, there would be multiple equilibria even if the sender has commitment devices, and intersubjective beliefs would best explain coercive success. If the target has a long time horizon, the sender has a short time horizon, and there are no commitment devices, only the no-concessions equilibrium would exist and the lack of commitment devices would predict coercive failure.

Note that the studying the interaction of time horizons, commitment devices, and intersubjective beliefs to explain coercive success essentially becomes a decision tree. First, the researcher should identify the time horizons of the relevant actors. Only then can the researcher determine whether focusing on commitment devices or intersubjective beliefs is more fruitful.

5.1 Time horizons

In general, the first step in explaining and predicting coercive success and failure is determining the time horizons of the two parties. If the time horizons are short, the primary determinant of coercive success will be the existence and form of commitment devices. If

the time horizons are long, the primary determinant of coercive success will be socially constructed intersubjective beliefs, and coercion will work when and only when both sides believe it will work.

Note that time horizons mean that the finite versions of the model are essentially special cased of the infinite horizon versions. When the time horizons are short, each infinite horizon model has only a single equilibrium, which matches the equilibrium in the corresponding finite models. For example, only the no-concessions equilibrium exists in both the finite and infinite horizon versions of the basic coercion model when time horizons are short. Thus, each infinite horizon model converges on the finite version of the model when time horizons are short.

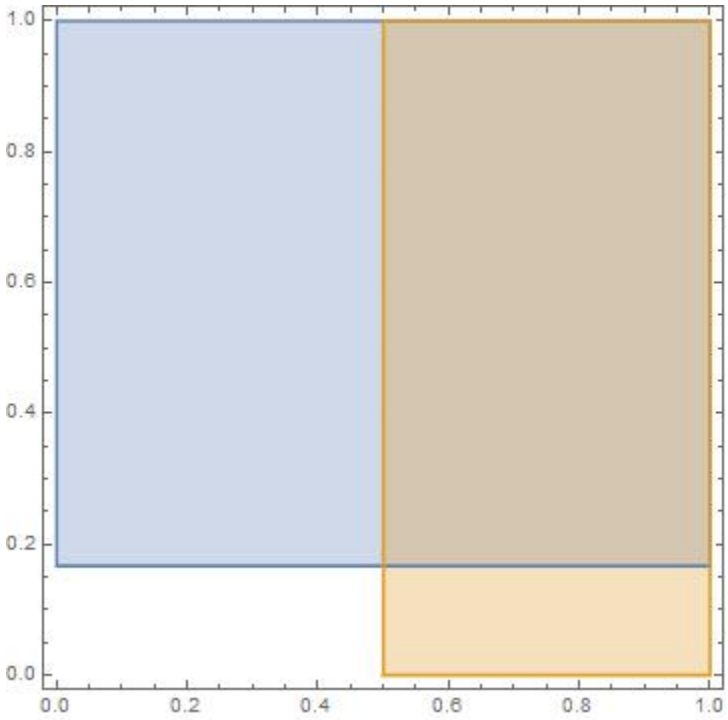
Substantively, the finite versions of the model also make sense as special cases of the infinite horizon models. International relations is inherently an ongoing game, and similar situations may repeat. However, it is possible that situations repeat so rarely that actors will not seriously consider future scenarios as the determine what to do in the present. In essence, when situations are repeated rarely the actors would have short time horizons. Thus, time horizons become the defining feature about whether there are multiple equilibria. Accordingly, time horizons are the defining feature about whether commitment devices or intersubjective beliefs is most relevant to explaining coercive success.

However, the form of the model may still matter in determining how long the time horizons need to be for the specific model form to no longer matter. In this paper, I have examined two extremes: when there is no commitment device and when the sender can irrevocably commit to punish if their demands are not met. It is likely, though not certain, that the needed discount parameters to create multiple equilibria are the highest in these models. Probably, other forms of commitment devices would need the same or lower discount parameters to create multiple equilibria.

The form of the model also matters if one actor has a relatively long time horizon while the other has a relatively short time horizon. In this case, the time horizons interact

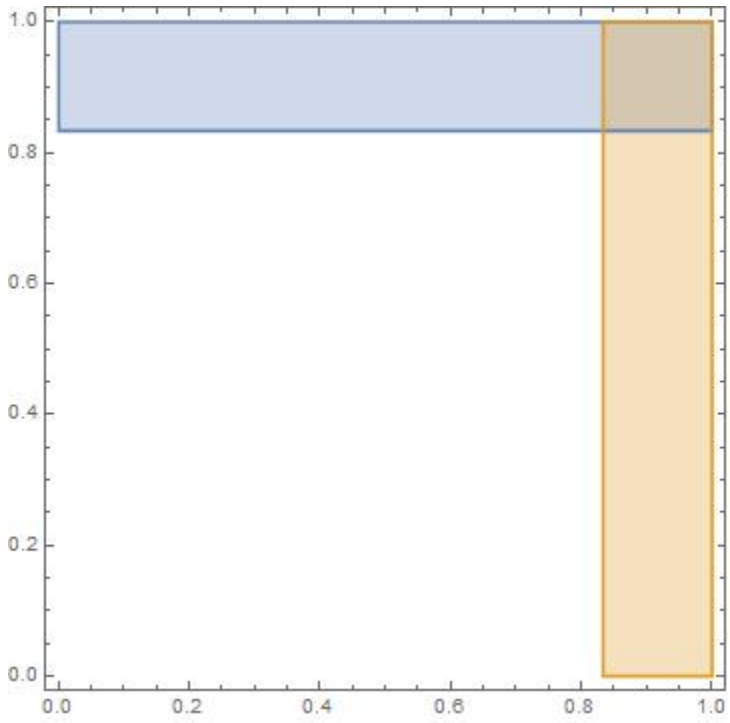
with the model form to determine whether there are multiple equilibria, and thus whether commitment devices or intersubjective beliefs better explain coercive success. In the basic coercion model, there are multiple equilibria whenever the sender has a long time horizon, regardless of the target's time horizon. Conversely, when the sender has a short time horizon, coercion is never successful regardless of the target's time horizon. In the tying hands model, there are multiple equilibria when the target has a long time horizon, but only a concessions equilibrium when the target has a short time horizon, regardless of the sender's time horizon.

Figure 3: Existence of Multiple Equilibria, costs below 1;
δ_T on horizontal axis, δ_S on vertical axis
$c_T = 0.5$, $c_S = 0.1$
Blue: Multiple equilibria in basic model, Tan: Multiple equilibria in tying hands model



Figures 3 and 4 show when there are multiple equilibria in both the basic model and the tying hands model. Where both ranges overlap, both models have multiple equilibria. Given that these are the extreme possibilities it is likely that multiple equilibria exist regardless of which commitment devices exist. Accordingly, within this range intersubjective beliefs will usually determine whether coercion is successful. Where neither range exists,

Figure 4: Existence of Multiple Equilibria, costs above 1;
δ_T on horizontal axis, δ_S on vertical axis
$c_T = 5$, $c_S = 5$
Blue: Multiple equilibria in basic model, Tan: Multiple equilibria in tying hands model



neither model has multiple equilibria. While some commitment devices may create multiple equilibria, in general the existence of commitment devices or lack thereof will determine coercive success. Where one model has multiple equilibria and the other does not, how to explain coercive success depends on whether commitment devices exist.

One interesting thing to note, is that the discount parameter the target needs to create a no-concessions equilibrium even with commitment devices depends only on the target's own costs. However, the discount parameter the sender needs to create a concessions equilibrium even if they do not have commitment devices may depend on the relationship between the sender's costs and the target's costs. This is because the maximum concessions the sender can get also depends on the target's costs.

5.2 Short time horizons: commitment devices

As described above, if both actors have short time horizons, whether coercion works depends on the existence and form of commitment devices. In addition, the length of time horizons needed to create multiple equilibria may also depend on which commitment devices exist. Here, I will explore further which commitment devices may exist that could influence coercive success.

Commitment devices are necessary to gain concessions when time horizons are short as it would be irrational for the sender to actually carry out their punishment threat. Punishment would also impose costs on the sender, but does not directly do anything to help the sender achieve their objectives. Knowing that the sender's threat is irrational, the target would simply refuse to make concessions. Thus, the sender needs some way to either ensure that the punishment threat would be carried out if the target fails to make suitable concessions, or some way for the sender to incentivize themselves to carry out the threat despite its costs. A few possibilities suggest themselves, although there may be additional commitment devices not listed here.

The first commitment device the sender may have is if they can create a situation where the target's own actions initiate punishment. For instance, a state may want to deter an invasion in which they know they will lose, but believe they can inflict significant costs before they lose. By deploying military forces on the border, combat and the associated costs becomes automatic if the target of deterrence does decide to invade. The ability to make punishment happen automatically is likely confined to deterrence situations, as it would be difficult to set up a situation where punishment will occur automatically if the the target fails to actively make concessions in a compellence scenario. However, making punishment automatic is not always possible even in deterrence situations. For instance, economic sanctions to deter human rights abuses have to be actively implemented, even if they are intended to deter rather than compel.

The brinksmanship scenarios common in theories of nuclear coercion are a variation

on making punishment occur automatically. Rather than punishment always occurring automatically if the target fails to make concessions, in brinksmanship scenarios the failure to make concessions creates some risk that punishment will occur automatically. An example would be NATO the deployment of tactical nuclear weapons in Europe. If the Soviet Union attacked, there was some possibility that the forces would go ahead and use these weapons without further orders, although this would not occur for certain. Since brinksmanship scenarios are simply a variant of making punishment occur automatically, they are also likely more common in deterrence scenarios.

A second related type of commitment device would be if the sender can create some legalistic framework separating who decides to implement punishment from who decides to make the threat and set up the commitment device. In this case, the leadership of the sender gives some conditional orders to subordinates to carry out punishment if and only if the target fails to make concessions. For this framework to actually serve as a commitment device, several conditions are necessary. First, the leadership must be able to give clear, conditional orders or parameters to a subordinate or another actor. Second, the subordinate must be able to judge whether the the target has made sufficient concessions, and thus whether to implement punishment. Third, the subordinate must be committed to carrying out the conditional orders as written, without exercising independent judgment about whether they are wise. In other words, the subordinate must be committed to carrying out the punishment action that they know is costly simply because their conditional orders demand it. Finally, the leadership must be unable to reverse the order between the time it becomes apparent that the target is not complying and the time the subordinate would initiate the punishment action.

Examples of this sort of framework include national judicial systems and gunboat diplomacy. In judicial systems, general laws are passed with appropriate punishment, and the implementation of those laws is left to prosecutors and judges. The prosecutors and judges, ideally, should be able to exercise independent judgment if the law has been broken,

but are committed to fulfilling their duty to implement the law without considering whether doing so is wise. Finally, it takes time to change the law, and so the executive or legislature may have difficulty changing the law between when a crime is committed and the trial and sentencing.

Gunboat diplomacy prior to the twentieth century also fulfilled these conditions.[10] The national leadership could issue coercive threats and issue orders to military commanders to implement some form of punishment if appropriate concessions were not made. However, because of the lengthy communications time, the national leadership could not countermand the orders easily. Thus, the local military commander had to decide whether to fulfill the punishment threat. Assuming they were committed to fulfilling their orders, the national leadership had in essence tied their hands on whether punishment would occur.

A third commitment device might involve hostage-taking or blackmail scenarios. In these cases, the sender has to pay the costs of punishment up front before the target makes concessions, to give themselves the opportunity to punish the target. However, once they have created the opportunity to punish the target, actually implementing the punishment is costless. Thus, the sender has no further disincentive to carrying out the punishment if the target fails to make concessions. However, a condition for hostage taking or blackmail to work is that carrying the punishment after the initial costs have been paid must be virtually costless. Even if the sender paid relatively minor costs in fulfilling the punishment threat after the target made their decision, it would be irrational to carry out the punishment.

A fourth possible commitment device would be issuing a public threat that would inflict audience costs on the sender if the target both failed to gain appropriate concessions and failed to punish. If the sender's audience costs outweighed their costs of punishing, it would no longer be irrational to carry out the threat. Since the sender would issue the threat before the target made their decision, it would also be rational to issue the threat in order to gain concessions. However, issuing the threat might also impose audience costs on

[10]I should note that gunboat diplomacy was often in pursuit of immoral ends. Here, I am considering simply whether it would work, and not whether its use was moral.

the target if they did make concessions, possibly affecting whether this strategy would be effective. There has been significant work on audience costs in political science , which could help determine when audience costs would serve as an effective commitment device.

While some commitment device is necessary to for coercion to be successful, the analysis cannot stop there. The target may themselves have commitment devices that they can use to commit or incentivize themselves not to make concessions. Thus, the actual outcome of a coercive interaction with short time horizons will depend on the interaction of the two sides commitment devices. If there are no commitment devices on either side, or only the target has commitment devices, coercion will inevitably fail. If only the sender has commitment devices, coercion may work, depending on the exact nature of the commitment device and associated parameter values. If both sides have commitment devices, coercion may or not work depending on the interaction of the two sides commitment devices and the parameter values.

One important, but not determinant factor, would be which side has the opportunity to publicly implement their commitment device first. If the sender goes first, coercion has a good chance of working. If the target implements their coercion device, coercion will likely fail. However, the exact outcome would depend on the specific nature of the commitment devices, how they interact, and how they are affected by parameter values such as the costs of punishment.

5.3 Long time horizons: norms and intersubjective beliefs

If the actors have long time horizons, commitment devices no longer become important in determining whether coercion will work. Coercive success is possible even without commitment devices, while coercive failure is possible even with commitment devices. However, given that the models contain multiple equilibria with long time horizons, explaining coercive success requires explaining which equilibrium is actually played. Equilibrium selection in turn requires reference to socially constructed intersubjective beliefs, as there is no ex-

ogenous reason one would occur over the other. Thus coercion will work when and only when both sides believe it will work. Here I will suggest some possible influences on these intersubjective beliefs, and thus on coercive success.

Since coercive success often depends on intersubjective beliefs that are socially constructed, it is likely that social norms are critical in determining coercive success. Norms are in essence intersubjective beliefs about what should happen, and thus would be closely related to the intersubjective beliefs about the outcome of the coercion game that determines when coercion succeeds. Indeed, Goertz (2003)[11] argued that equilibria in formal models are in fact a type of norm. Thus, looking at which norms exist and their formation should provide insight into which equilibria is likely to occur, and thus whether coercion will be successful. Three norms in particular stand out: norms related to preserving the status quo, norms related to hierarchy, and norms about the legitimate distribution of the disputed good.

First, it is likely that coercion is more likely to succeed when it attempts to preserve a percieved status quo rather than alter the status quo. There are fairly clear international norms against using threats to alter the status quo (e.g. Zacher 2001; Fazal 2004;). However, international norms generally state that defending the status quo is acceptable . In addition, there are psychological biases towards preserving the status quo (e.g. Kahneman, Knetsch, and Thaler 1991). Thus, both sides are more likely to believe that coercion will work when the coercive threat is intended to preserve the percieved status quo rather than alter the status quo. This restores the distinction between deterrence and compellence, although due to norms rather than any material factors. Note in addition, that what is important is the percieved status quo rather than the de facto status quo.

Second, I believe coercion is more likely to work when the threat sender has perceived authority over or is otherwise above the target in a hierarchy. While international relations appears anarchic, Lake (2009) has shown that there are a number of informal hierarchies.

[11] International Norms and Decisionmaking - Check citation - I think this is it, but not sure

Within these hierarchies, there is a belief that the subordinate actor should generally follow the superior actor's lead and direction. Thus, when the threat sender is the superior actor, the subordinate target would be more likely to believe that coercion would work, which would make it actually more likely to work.

Similarly, when the sender's demand is backed up by international organizations, coercion would also be more likely to work. While international organizations have little power to inflict punishment themselves, they can authorize other states or actors to inflict punishment. In addition, international organizations would be perceived to have some authority over there members, at least within the organizations mandate. Thus, if an international organization backs the sender's demand, the immediate actors are more likely to believe that the sender's coercive threat will work, and coercion will be more likely to work. Conversely, if the organization backs the target, coercion will be more likely to fail for the same reasons.

The third likely influence on whether coercion is successful is international laws or norms dictating the distribution of the particular disputed issues. There are numerous guidelines for who deserves what in disputes, ranging from formal treaties to customary law and norms. These rules create beliefs about what the actors should do. Accordingly, if these rules suggest that the threat sender deserves to be given the disputed issue or that the target otherwise should follow the sender's demand, both sides will be more likely to believe that the target will give in, and coercion will become more likely to work. For instance in maritime disputes, if the target has violated the UN Convention on the Law of the Sea (UNCLOS), the sender's threats to get the target to comply with UNCLOS would be more likely to work. In contrast, if the target is attempting to get the sender to do something contrary to UNCLOS, the threats are less likely to work.

Note that in all three influences on equilibrium selection work through influencing the actors beliefs about whether coercion will work. This has a couple of significant implications. First, this is not an exhaustive list of what factors could influence intersubjective beliefs about whether coercion would be successful. There are likely other factors. It is even possible that

the factors I identified are relatively minor influences on coercive success, although they do appear the most obvious influences.

Second, even if these influences on coercive success generally hold, they would not always have consistent effects on whether coercion generally holds. It is certainly possible for other factors to have a greater influence on the beliefs about whether coercion will work, negating the influence of these factors individually or collectively. In addition, it is possible that the actors will develop collective beliefs contrary to what these factors indicate for idiosyncratic reasons. Thus, overall, further research is needed to understand how states and other actors form their beliefs about whether coercion will work in order to fully explain coercive success and failure.

6 Conclusion

In this paper, I have reanalyzed several basic models of coercion by punishment. I analyzed a basic model, in which the sender threatens the target to gain concessions. I then analyzed a model in which the sender could tie their hands, such that punishment would occur automatically if the target failed to make sufficient concessions. I began by analyzing each model in a single-stage or finite format. Then, to account for the fact that coercive scenarios will not happen in isolation, I reanalyzed each model in an infinite horizon format.

A basic problem across all of these models is maintaining the credibility of the coercive threat. Since carrying out the punishment would likely be costly to the sender, absent some commitment device or other motivation, it would always be irrational for the sender to carry out the punishment threat when the target failed to make concessions. Thus, the problem of credibility is not limited to nuclear conflict, but extends to all coercive scenarios. The need to be able to credibly commit to inflict punishment fundamentally shapes the explanations for coercive success across all the models.

However, the means by which the sender can credibly commit, and thus the expla-

nations for coercive success and failure, vary fundamentally according to the actors' time horizons. If the time horizons are short, coercive success is explained by the existence or absence of any commitment devices and the form those commitment devices take. However, if the time horizons are long, the commitment devices become unimportant, and coercive success is best explained by socially constructed intersubjective beliefs. Thus, time horizons are not simply one factor that affects coercive success, but fundamentally reshape the best explanations for coercive success. The centrality of time horizons does not appear to be recognized by previous work.

If time horizons are short, whether coercion works is determined primarily by the existence and form of commitment devices. The sender needs some way to tie their hands or otherwise incentivize themselves to carry out the punishment threat. These commitment devices can range from the ability to make punishment self-executing, to legalistic mechanisms separating the decision to issue a coercive threat from the people who carry out punishment, to audience costs. At the same time, the target may have their own commitment devices that would counteract the sender's ability to commit. While the importance of commitment devices has been explored before, future research is still needed to further develop the types of commitment devices and their impact on coercive success.

If time horizons are long, in contrast, coercive success is explained primarily by socially constructed intersubjective beliefs. With long enough time horizons, both models have multiple equilibria: some in which coercion works and some in which coercion fails. These occur because the sender can become willing to punish in order to secure future concessions, while the target may resist making concessions even if they believe they will be punished in order to avoid concessions in the future. These reputational impacts override the existence or non-existence of commitment devices. Equilibrium selection cannot be explained by material factors with long enough time horizons. Thus, the equilibrium that actually occurs must be explained by referring to socially constructed intersubjective beliefs. In essence, coercion will work when and only when both sides believe it will work. These beliefs might

be impacted by status quo biases, international norms, and perceptions of hierarchy. Future research can further develop and test which factors impact the beleifs about coercive success.

While a few papers have considered whether social constructivism might impact coercion, the need to refer to socially constructed factors to explain coercive success does not appear to have been realized. This finding likely also has broader implications for international relations theory. Coercion is an form of hard power, and thus should be a less likely case for socially constructed factors to play a significant role. The finding that socially constructed factors may nevertheless determine whether coercion is successful likely has implications for the broader relationship between socially constructed factors and material factors in explaining international relations.

References

Abrahms, Max. 2011. "Does Terrorism Really Work? Evolution in the Conventional Wisdom since 9/11." *Defense and Peace Economics.* 6: 583-594.

Anderson, Nicholas D., Alexandre Debs, and Nuno P. Monteiro. 2019. "General Nuclear Compellence: The State, Allies, and Adversaries." *Strategic Studies Quarterly.* 13(3): 93-121.

Beardsley, Kyle and Victor Asal. 2009. "Winning with the Bomb." *Journal of Conflict Resolution.* 53(2): 278-301.

Danilovic, Vesna. 2001. "Conceptual and Selection Bias Issues in Deterrence." *The Journal of Conflict Resolution.* 45(1): 97-125.

Drezner, Daniel W. 1998. "Conflict Expectations and the Paradox of Economic Coercion." *International Studies Quarterly.* 42(4): 709-731.

Fazal, Tanisa M. 2004. "State Death in the International System." *International Organization.* 58(2): 311-344.

Fearon, James D. 2002. "Selection Effects and Deterrence." *International Interactions.* 28: 5-29.

George, Alexander L. 1991. *Forceful Persuasion: Coercive Diplomacy as an Alernative to War.* Washington D.C.: United State Institute of Peace Press.

Goertz, Gary. 2003. *International norms and decision making : a punctuated equilibrium model.* Lanham, Maryland: Rowman & Littlefield.

Jervis, Robert. 1979. "Deterrence Theory Revisited." *World Politics.* 31(2): 289-324.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1991. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias." *The Journal of Economic Perspectives.* 5(1): 193-206.

Kilgore, D. Marc and Frank C. Zagare. 1991. "Credibility, Uncertainty, and Deterrence." *American Journal of Political Science.* 35(2): 305-334.

Kroenig, Matthew. 2013. "Nuclear Superiorty and the Balance of Resolve: Explaining Nuclear Crisis Outcomes." *International Organization.* 67(1): 141-171.

Krustev, Valentin L. 2010. "Strategic Demands, Credible Threats, and Economic Coercion Outcomes." *International Studies Quarterly.* 54(1): 147-174.

Kydd, Andrew H. and Roseanne W. McManus. 2015. "Threats and Assurances in Crisis Bargaining." *Journal of Conflict Resolution.* 61(2): 325-348.

Lacy, Dean and Emerson M.S. Niou. 2004. " A Theory of Economic Sanctions and Issue Linkage: The Roles of Preferences, Information, and Threats." *The Journal of Politics.* 66(1): 25-42.

Lake, David A. 2009. *Hierarchy in International Relations.* Ithaca, New York: Cornell University Press.

Lebow, Richard Ned and Janice Gross Stein. 1989. "Rational Deterrence Theory: I Think, Therefore I Deter." *World Politics.* 41(2): 208-224.

Miller, Nicholas L. 2004. "The Secret Success of Nonproliferation Sanctions." *International Organization.* 68(4): 913-944.

Moller, Sara Bjerg. 2013. "So Which is It? Deterrence or Compellence?" *Political Violence at a Glance.* August 28, 2013. https://politicalviolenceataglance.org/2013/08/28/so-which-is-it-deterrence-or-compellence/

Nash, John. 1951. "Non-Cooperative Games." *Annals of Mathematics.* Second Series. 54(2): 286-295.

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

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

Powell, Robert. 2015. *Nuclear Brinkmanship, Limited War, and Military Power. International Organization.* 69(3): 589-626.

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

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

Sechser, Todd S. and Matthew Fuhrmann. 2013. "Crisis Bargaining and Nuclear Blackmail." *International Organization.* 67(1): 173-195.

Signorino, Curtis S. and Ahmer Tarar. 2006. "A Unified Theory and Test of Extended Immediate Deterrence." *American Journal of Political Science.* 50(3): 586-605.

Slantchev, Branislav. 2003. "The Power to Hurt: Costly Conflict with Completely Informed States." *American Political Science Review.* 47(1): 123-135.

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

Smith, Alastair. 1996. "The Success and Use of Economic Sanctions." *International Interactions.* 21(3): 229-245.

Snyder, Glenn H. 1960. "Deterrence and Power." *The Journal of Conflict Resolution.* 4(2): 163-178.

Trager, Robert F. 2013. "How the scope of a demand conveys resolve." *International Theory.* 5(3): 414-445.

Trager, Robert F. and Dessislava P. Zagorcheva. 2006. "Deterring Terrorism: It Can Be Done." *International Security.* 30(3): 87-123.

Tsebelis, Gearge. 1990. "Are Sanctions Effective? A Game-Theoretic Analysis." *The Journal of Conflict Resolution.* 34(1): 3-28.

Whang, Taehee, Elena V. McLean, and Douglas W. Kuberski. 2013. "Coercion, Information, and the Success of Sanction Threats." *American Journal of Political Science.* 57(1): 65-81.

Zacher, Mark W. 2001. "The Territorial Integrity Norm: International Boundaries and the Use of Force." *International Organization.* 55(2): 215-250.

Zagare, Frank C. 1990. "Rationality and Deterrence." *World Politics.* 42(2): 238-260.