International Interactions, 28:5–29, 2002

0305-0629/02 $12.00 + .00

SELECTION EFFECTS AND DETERRENCE

JAMES D. FEARON

Department of Political Science, Stanford University, Stanford, California, USA

(Received for Publication September 7, 2001)

The empirical question of how often deterrent threats issued during international disputes succeed has been hotly debated for years, with some researchers arguing that virtually no robust cases of success can be identified. I argue that what appears to be an empirical and methodological debate actually arises from the inadequacy of classical rational deterrence theory, which fails to comprehend the implications of states' strategic self-selection into international disputes. Rational self-selection is shown to imply that in a sample of crises, deterrent threats issued after an initial challenge will tend to fail in precisely those cases where they are relatively most credible signals of an intent to resist with force. The product of a *selection effect*, this paradoxical implication allows a resolution of the debate on the efficacy of deterrence in crises. And because selection effects can arise whenever a historical "case" is the product of choices by actors who also influence the outcome in question, this example from the study of deterrence has broad relevance for empirical research.

KEY WORDS: international crisis bargaining, rational deterrence theory, selection effects.

INTRODUCTION

How often and under what conditions do threats issued in the course of an international dispute successfully deter aggressive action by the state challenging the status quo? This important and apparently straightforward empirical question pro-

---

This article draws on chapter 5 of James D. Fearon, "Threats to Use Force: Costly Signals and Bargaining in International Crises," Ph.D. dissertation, U.C. Berkeley, 1992, and was originally intended for publication in an edited volume that never materialized. For helpful comments I wish to thank Chaim Kaufmann, Andrew Moravscik, Bruce Russett, and especially Charles Glaser. Some of the research reported here was funded by the Institute on Global Conflict and Cooperation.

Address correspondence to James D. Fearon, Department of Political Science, Stanford University, Stanford, CA 94305-2044, USA

voked a heated debate in the field of international relations in the late 80s and early 90s. On one side, Paul Huth and Bruce Russett argued that deterrent threats succeeded in 34, or almost 60 percent, of the 58 "extended immediate deterrence" crises they identified. On the other side, Richard Ned Lebow and Janice Gross Stein forcefully disputed this assessment. By their reading, only 10 of Huth and Russett's cases were properly regarded as "deterrence encounters" and at most two of these contained instances of successful deterrent threats.[1]

What made these dramatically different results surprising and puzzling was that the researchers agreed on the definition of extended immediate deterrence and deterrence success implied by rational deterrence theory. As Lebow and Stein noted, "We agree that immediate extended deterrence occurs only when an attacker contemplates military action against another country and a third party [the defender] commits itself to the defense of the country threatened with attack." Immediate deterrence succeeds when the challenging state fails to attack or undertake "a proscribed action" because it has been persuaded by the defender that the consequences would unacceptable (Lebow and Stein, 1990, pp. 242, 245; Huth, 1988, pp. 23–27; Morgan, 1977).

Given this agreement on definition, the participants explained their conflicting findings primarily as the result of disagreement on "the application of that definition to the cases" (Lebow and Stein, 1990a, p. 342).[2] The dispute and the puzzle thus appeared to be matters of coding—how to go about interpreting what happened in particular cases and whether to designate them as deterrence successes, failures, or not as deterrence encounters at all. The debate between Huth and Russett and Lebow and Stein is thus easily read as a product of the methodological divide between "large-N" statistical researchers who want simple, operational coding rules and "small-N" case study advocates who want detailed, textured interpretations of particular cases.

This reading is mistaken. Instead, what looks to be an empirical and methodological dispute over coding is better understood as reflecting the inadequacy of classical rational deterrence theory, the framework that structures both sides' approach to the data. Classical rational deterrence theory predicts that a state's threats or commitments will be more likely to deter aggression when they are more credible, meaning that a potential aggressor believes the threats are more likely to be carried out. The classical theory, however, focuses on a challenging state's decision to attack, while failing to consider the effects of strategic behavior by states prior to this decision, during crisis bargaining. I argue that signaling by states prior to the challenger's military decision will have a surprising consequence. In a sample of cases, we should find that immediate deterrent threats—that is, threats issued after an initial challenge—are most likely to fail when they are relatively *most* credible as indicators of the defender's intent to resist.

This is the result of a *selection effect.* Rational challengers select themselves into crises according to their prior beliefs about the defender's willingness to resist with force. To the extent that this occurs, the crises in which the defenders' immediate deterrent threats are most credible will tend to be crises in which the challenging states are relatively strongly motivated to change the status quo, and thus willing to accept an appreciable risk of conflict. Hence despite their greater credibility compared to immediate deterrent threats in other cases, defender threats in this subset are

less likely to succeed.

This observation allows us to make better sense of the empirical debate on immediate deterrence. If the selection effect operates, then excluding cases in which challenger or defender threats were relatively incredible should actually increase the rate of immediate deterrence failure in the remaining sample.

This, in essence, is what Lebow and Stein did with Huth and Russett's data set. On the grounds that rational deterrence theory requires credible threats in order for the theory to apply, Lebow and Stein argued for the exclusion of many cases in which the threats made were not very credible indicators of challenger or defender intentions. The results are as expected if states act strategically in crisis bargaining: Immediate deterrence fails more often in cases where the threats made were relatively credible.

From this perspective, the debate on extended immediate deterrence grows out of a gap in classical rational deterrence arguments. The classical theory does not distinguish between the deterrence of an initial challenge (general deterrence) and deterrence once a challenge has been made (immediate deterrence), and does not see that if states act strategically these will be quite different phenomena.[3] Taking their cue from the classical theory, both pairs of researchers in the debate saw immediate deterrence as an instance of the larger category deterrence, even while recognizing the general-versus-immediate distinction. Both sides assumed that evidence on immediate deterrent threats is directly relevant to assessing key propositions of the classical theory, such as the proposition that more credible threats are more likely to deter. But if selection effects operate, then immediate deterrence encounters cannot be treated as an empirically testable version of general deterrence, or any notion of deterrence in the abstract. Selection effects introduce systematic bias, so that relationships that may be true for general deterrence will appear exactly reversed for immediate deterrence.

This point has wider significance. Selection effects matter not just in international disputes but in a broad range of phenomena studied by scholars of comparative and international politics. The historical cases we use to evaluate theories—be they cases of military threats, foreign policy success, war, revolution, democracy, or whatever—typically became cases by virtue of prior choices made by individuals. Selection effects occur when factors that in uence the choices that produce cases also influence the outcome or dependent variable for each case. Failure to conceptualize the choice process that generates the cases can then yield incorrect inferences about what explains variation across cases. Seen in this light, the debate on the effects of threats used in international crises provides a nice example of a more general problem. Selection effects often matter but are infrequently seen or adequately understood.[4]

Thinking carefully about the choice process that generates a case can also help to refine a theory. In the present instance, I will argue that classical rational deterrence theory's failure to analyze the bargaining that precedes a state's decision to attack leads the theory to misapprehend what goes on in historical cases. Deterrent threats used in crises do more, and are meant to do more, than simply dissuade challengers from military attack. Rather, they are signals that communicate (in a noisy way) what bargains a defending state would or would not accept in preference to war—deterring potential attack is bound up with signaling preferences over possible nego-

tiated outcomes. Classical rational deterrence theory neglects this interplay between military threats and bargaining over a range of issue resolutions. I argue that in consequence the classical theory cannot generate the clear and unambiguous coding criteria required to test the theory, and that this problem plays an important but hidden role in the empirical debate on immediate deterrence.

The first section below briefly reviews what is at stake in the debate and considers the argument over coding standards more carefully. The second develops the argument on selection effects, showing how these arise from states' use of costly signals in crisis bargaining. The third shows evidence of selection effects at work in the data set by focusing on several cases from it. The fourth section considers classical rational deterrence theory's neglect of bargaining, and the implications for empirical tests of the theory. The fifth draws conclusions.

THE DEBATE ON EXTENDED IMMEDIATE DETERRENCE

The debate over how to interpret and code cases in the Huth and Russett data set has both practical and theoretical significance. On the practical side, extended immediate deterrence crises have figured prominently in the foreign relations of the U. S. and other great powers. As the data set indicates, key crises during the Cold War, before World Wars I and II, during 19th century colonial rivalry, and decolonization in the 50s and 60s have all involved issues of extended deterrence.

The problem of extended deterrence has not disappeared with the end of the Cold War. The initial rationale for the recent U.S. deployment in the Middle East in 1990 was to extend deterrence to Saudi Arabia. Arguably, the subsequent war with Iraq grew from a failure to issue a significant immediate deterrent threat on behalf of Kuwait.[5] Outside the Middle East, the question of an extended deterrent relationship between the U.S. and Taiwan is at the heart of U.S.–Chinese security relations. Thus, empirical work on the conditions under which extended immediate deterrent threats succeed or fail remains highly relevant.

On the theoretical side, both pairs of researchers naturally expected that good data on extended immediate deterrence would allow an empirical assessment of deterrence theory. Huth and Russett used their data to evaluate several hypotheses on "what makes deterrence work," most of which derived from classical rational deterrence theory. Thus, Huth hypothesized that "the probability of deterrence success" is greater the more the defender is favored by the balance of military capabilities or by the balance of interests, on the grounds that both factors influence the defender's credibility (Huth, 1988, pp. 35–47). Overall, Huth and Russett read their evidence as supporting rational deterrence theory. They found that even if they granted various of Lebow and Stein's concerns about coding, their data showed that "the overall predictive power of a fairly parsimonious rational model remains respectable" (Huth and Russett, 1990, p. 492).

This view conflicts with that of Lebow and Stein, who vigorously criticized rational deterrence theory as lacking any empirical foundation (Lebow and Stein, 1987; Lebow and Stein, 1989). In their own work they found that immediate deterrent threats very rarely succeeded (Lebow, 1981; Stein, 1985). Moreover, Lebow and Stein claim that this low rate of success could not be explained by rational deterrence

theory, but only by the operation of psychological pathologies in crisis decision-making. A challenging state's strong motivation to threaten makes its leaders subject to specific biases, and these lead them to ignore or react violently to the defender's counterthreats.

Lebow and Stein's interpretation of the cases in Huth and Russett's data set strongly supported their views on the empirical weakness of rational deterrence theory. Their reading of the historical evidence led them to discard 53 of the 67 cases they examined from Huth and Russett's 1984 and 1988 data sets. Of the 10 cases which they agree qualify for inclusion,[6] they originally designated eight as immediate deterrence failures and three as successes.[7] For two of these three cases (Munich 1938 and the Egyptian attempt to deter Israeli attack on Syria in 1967), they note that the success of immediate deterrence was "partial or short-lived"—"They could be classified as successes only by the most liberal application of our coding criteria." Consistent with their earlier research, Lebow and Stein (1990, p. 348) conclude that

> Among the most important findings with respect to the dependent variable is the seemingly elusive and fragile nature of the success of immediate deterrence. . . . Examination of these cases suggests that immediate extended deterrence successes tend to be uncommon, partial, and tenuous.

These words actually understate how remarkable is Lebow and Stein's empirical claim. They claim that an extensive review of international disputes in the last 90 years reveals only one robust case of extended immediate deterrence success—a U.S. threat to Turkey not to invade Cyprus in 1964. Moreover, as Huth and Russett note in their reply, this one case is not a case of extended immediate deterrence as they define it, since the U.S. threat was to suspend economic aid to Turkey, not to respond militarily (Huth and Russett, 1990, p. 474).[8] So if Lebow and Stein's reading of the evidence is valid, then they have helped discover one of the strongest interesting empirical regularities in international politics, on par with the observation that democracies almost never fight one another.

In addition, if they are right, Lebow and Stein appear to have found stronger empirical support for their argument that leaders in crisis bargaining are systematically affected by psychological biases that make them unable to form rational beliefs about the behavior of an adversary. Lebow and Stein make much of the fact that their empirical conclusions are based on operational definitions that are not "arbitrary," but are "derived directly from the fundamental axioms of deterrence theory" (Lebow and Stein, 1990a, p. 324). In particular, they require that the challenger's and defender's threats be credible as indicators of their intentions. This seems in line with rational deterrence theory's emphasis on the importance of the credibility of the defender's threat for successful dissuasion.

What accounts for Lebow and Stein's strikingly different appraisal of the same historical events? As noted, the two pairs of researchers basically agree on the definition of extended immediate deterrence; they differ in their "application of that definition to the cases" (Lebow and Stein, 1990a, p. 324). Lebow and Stein argue, and Huth and Russett concur (Huth and Russett 1990, p. 481), that a key difference between them is the amount and nature of the historical evidence they require to

conclude that the attacker intended to attack, and the defender intended to defend.

Huth and Russett rely on behavioral indicators: Did the challenger and defender make statements indicating intent to attack or defend? Did they mobilize and position troops in a manner consistent with such intentions? For example, for challengers, the operational criteria were that "at a minimum, one or more of the political and military elites that are shaping the foreign policy behavior of the attacker country must be recommending or considering using military force, and that this behavior leave observable traces in the form of actions or statements in the name of the state" (Huth and Russett, 1990, p. 482). Lebow and Stein argue that these behavioral criteria are not reliable indicators of states' intentions. Because "Military deployments . . . can be used for a wide range of purposes," "intention can only be established by reference to other kinds of historical evidence" (Lebow and Stein, 1990a, p. 342). Although Lebow and Stein are somewhat vague about just what evidence they require—they refer to "multiple streams of evidence interpreted in context"—they seem to put a premium on documents and their own reconstructions of what they find plausible. Most of the 41 cases that Lebow and Stein discard are thrown out because they were, in their words, "not deterrence encounters:" Lebow and Stein find insufficient evidence to conclude that either the challenger, the defender, or both "seriously intended" to attack or defend.

In their reply, Huth and Russett accept that "Lebow and Stein are correct to focus attention on [the] important problem" of establishing intent (Huth and Russett, 1990, p. 481).[9] They argue, however, that Lebow and Stein's "emphasis on using documentary evidence and the consensus of diplomatic historians and country experts to determine the true intentions of the attacker is plagued with methodological difficulties." Huth and Russett "deliberately refrained" from assessing how serious was the attacker about using force because "[t]o do so in accordance with the standards of good social science would require that a set of operational rules and guidelines (that are independent of the known outcomes) be formulated and then consistently applied." They note correctly that Lebow and Stein have not done this (1990, p. 482).

The debate thus seems to center on the choice of methodological standards for evaluating state leaders' intentions. If we adopt Lebow and Stein's methodological criteria, we get empirical results that appear to favor psychological over rationalist theories of deterrence. If we adopt Huth and Russett's criteria, rational deterrence theory appears supported.

COSTLY SIGNALING AND SELECTION EFFECTS IN CRISIS BARGAINING

Classical rational deterrence theory is inappropriate for structuring the empirical analysis of immediate deterrence. Amounting to an expected utility analysis of a state's decision to attack, the classical theory does not analyze the strategic interaction that characterizes states' efforts to signal their intentions or commitments. In consequence it misunderstands (or is misapplied to) the problem defenders face by the time immediate deterrence is at issue. I begin by briefly characterizing the classical model; then I sketch an empirically defensible alternative that incorporates signaling; and finally I consider how rational states would act if faced with this strategic

situation.

In its simplest form, classical rational deterrence theory focuses on two decisions—a potential challenger's decision whether to attack (or take some aggressive action), and a potential defender's decision whether to fight if an attack occurs. Figure 1 represents this sequence, which appears often in classical rational deterrence theory (Ellsberg, 1975; Snyder, 1961; Kaufmann, 1954; Achen and Snidal, 1989).

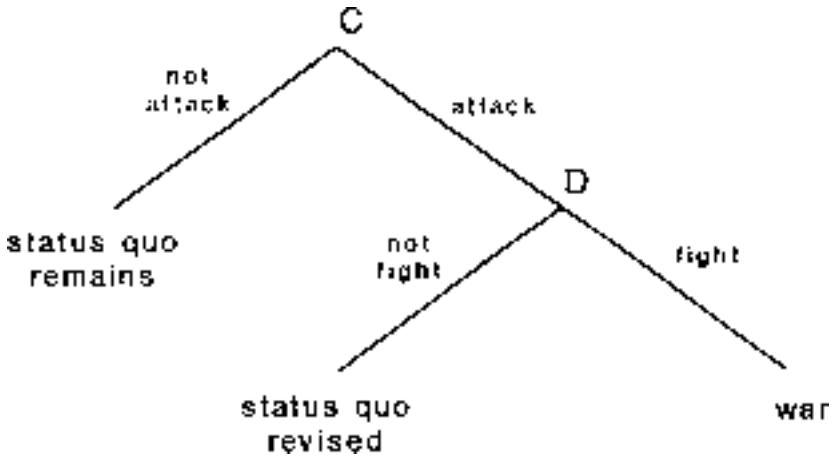

Figure 1. The classical rational deterrence model.

The classical theory's analysis of the problem yields the following common-sense proposition: Other things equal, a challenger will be less likely to attack the greater the expectation that the defender would choose to offer resistance if an attack occurs. The defender's threat to resist aggression is said to be more "credible" the greater the perceived probability that it would be carried out.[10]

On the basis of this argument about credibility, the classical theory proceeds to draw inferences about the effects of signaling by the defender prior to the challenger's choice in Figure 1. The core inference is that successful efforts by the defender to signal the credibility of its threat to resist will reduce the likelihood of a challenge. This inference has been used to support the following general hypothesis. *If a sample of cases is divided according to some measure of defender credibility, we should see fewer "deterrence failures" in the set with more credible threats.* Almost all of Huth and Russett's specific hypotheses derive from this core proposition.

While the hypothesis may well be true for general deterrence, I argue below that the reverse should hold for immediate deterrence if state leaders act strategically. The classical theory draws an inference about the effect of prior signaling by the defender without explicitly analyzing its role in crises. When we carry out this analysis using a game-theoretic model of crisis bargaining, we get a different and richer understanding of the deterrence problem imbedded in it.

Immediate deterrence becomes an issue only if general deterrence fails, that is, if

a state chooses to take an action indicating that it might attack or make some other undesired move in the near future. Initial challenges may include threats or demands issued by state leaders, the mobilization and deployment of troops, or a unilateral effort to change the status quo. Confronted with an initial challenge, the defending state next has a choice about whether to try immediate deterrence—in other words, to signal its willingness to resist with force if the challenging state acts or continues to act on its initial threat. Both choices precede the decisions considered explicitly in the classical rational deterrence model of Figure 1. Both are signals used by states in order to communicate their intentions and to test the other side's, prior to deciding whether to use of force. Figure 2 explicitly represents these prior signals together with the subsequent choices examined by the classical theory.

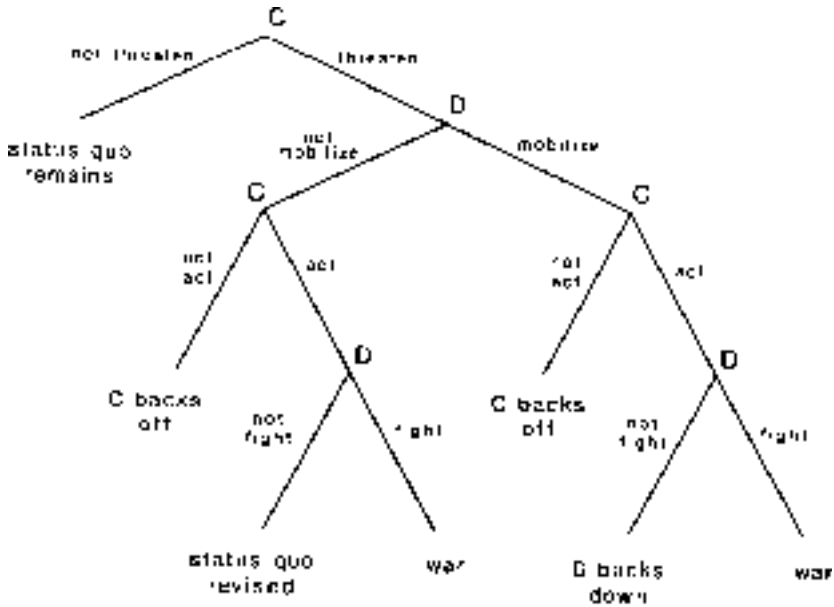

Figure 2. A simple model of an immediate deterrence crisis. Note: "mobilize" refers to any immediate deterrent effort by the defender.

While it is far simpler than any particular historical case, the crisis model of Figure 2 has an empirical foundation. For example, empirical work on immediate deterrence has implicitly used the model to represent international crises. To qualify for inclusion in Huth and Russett's data set, a case must include an observable initial threat by a challenging state, followed by an observable counterthreat by the defending state. Thus, in the model's terms, Huth and Russett select cases that reach the challenger's choice after observing an immediate deterrent response by the defending state. They then ask when challengers tend to choose "act" (immediate deterrence failure) rather than "not act" (immediate deterrent success), and, in their 1988

article, about the defender's decision to fight if the challenger acts.[11]

Despite their criticism of Huth and Russett's coding, Lebow and Stein employ essentially the same underlying model to describe how international crises unfold empirically. Lebow and Stein characterize crises as sequences of action and reaction, beginning with an Initiator's decision about whether to challenge a Defender, followed by the Defender's choice of whether to "reinforce deterrence," followed by the Initiator's choice of whether to pursue its challenge or to let it drop (Lebow and Stein, 1990b, p. 55). This is again the pattern represented by Figure 2.[12]

How would rationally-led states act given this set of choices? The problem is more complicated and interesting than that considered by the classical model, since states must now worry about sending and interpreting signals. For example, suppose the challenger threatens, expecting that the defender is unlikely to be willing to fight over the issue. What should it conclude about the defender's resolve if the defender actually *does* try a counterthreat such as partial mobilization? That the defender is likely to be bluffng, or that resistance is probable? Part of the difficulty is that the defender's incentives to bluff, which influence its credibility, may depend on its assessment of the credibility of the challenger's initial threat. If the challenger threatened hoping for "cheap gains," an immediate deterrent response might be worth trying even if the defender was reluctant to fight. But in turn, the credibility of the challenger's initial threat (as an indicator of its determination to act) may depend on its assessment of the defender's likely response. To understand the logic of this strategic interdependence and the logic of the problem, we require a game-theoretic analysis that allows for states' uncertainty about their opponents' preferences.

Carried out elsewhere, this analysis yields two broad conclusions about how crisis signaling works that are relevant here (Fearon, 1990; Fearon, 1992). First, signaling will allow states to credibly reveal their actual willingness to attack or to resist with force only if the signals are costly in a specific way: Their expected cost must be greater for a state with a low value for conflict (an opportunist) than for one with a high value (a motivated state). In terms of Figure 2, the challenger's initial threat will convey information about its willingness to act only if it creates costs that would be suffered by the state's leaders if they were to back down after the defender attempted immediate deterrence. Likewise, an immediate deterrent threat should rationally lead the challenger to revise its initial beliefs about the defender's credibility only if it creates costs the defender would suffer if the challenger acted and the defender then backed down.[13]

Typically, audience costs serve to make states' signals informative in crises. Leaders face domestic political audiences concerned with whether the country's foreign policy is successful. If backing down after having made a show of force exposes them to more serious domestic political criticism than would quietly ceding the issues at stake, a threat can rationally have some effect on another state's beliefs about one's intentions and credibility.[14]

The second conclusion follows from the first. If crises are, in effect, sequences of costly signals, then selection effects will operate. States select themselves into and out of crises according to their beliefs about an opponent's willingness to use force and their own value for conflict on the issues at stake.[15] With costly signaling, opportunistic challengers or defenders are more likely to drop out at each stage of the crisis

than are more motivated states. Thus, as a crisis proceeds it becomes increasingly likely that the states involved both have high values for conflict on the issues in dispute. In this view, an international crisis acts something like a filter that gradually separates out more highly motivated states.

How exactly does this work? If it is costly for a leader to make a threat and then back down if resistance is met, a state with relatively low resolve will be less likely to challenge general deterrence than will a state with higher resolve.[16] The reason is that a low resolve state is more likely to wind up paying the audience costs of backing down if the defender chooses to resist. Thus, leaders who choose to challenge general deterrence will have, on average, higher levels of resolve (or motivation) than will leaders who are merely considering a challenge to general deterrence. This is a selection effect—if threats are costly signals, then on average relatively motivated states will choose to threaten.[17]

What determines how strongly motivated, and thus how amenable to immediate deterrence, a state that chooses to challenge will be? Though many factors are involved, the game-theoretic analysis points in particular to prior beliefs—that is, to precrisis expectations about the likelihood that the other side would be willing to fight. Prior beliefs about the defender strongly influence what sort of state would choose to challenge initially, and this fact has strategic consequences that ramify throughout a crisis.[18]

Consider the initial decision to challenge. The greater the challenger's precrisis belief the defender would be willing to resist with force, the less likely a challenge, since an opportunist is more likely to be deterred by the prospect of paying audience costs. Thus general deterrence is stronger.

However, if a challenge occurs—general deterrence fails—the challenger is more likely to be highly motivated the greater was the prior belief that the defender would resist. If challengers rationally select themselves into crises, then the greater the prior expectation that the defender might fight, the more strongly motivated must a challenger be to assume the risks of a challenge. It follows that an initial threat will be a more informative and credible indicator of the challenger's intention to attack, the greater was the challenger's initial expectation of resistance.

In turn, this initial selection into a crisis by the challenger has a strategic consequence for the defender's immediate deterrent threat: the credibility of the defender's response will tend to vary with the credibility of the challenge. The defender has less incentive to bluff—to try immediate deterrence if not actually willing to fight—the greater its belief that the challenger is highly motivated and not an opportunist seeking cheap gains. So the defender's immediate deterrent threat will be a more credible indicator of its willingness to resist, the more credible the initial challenge.

The same reasoning applies in the other direction. The less the defender is initially expected to be willing to resist a threat or demand, the greater the incentive for opportunistic challenges. Thus the challenger's initial threat will be relatively *in*credible the less the challenger expected a tough response from the defender. In turn, an immediate deterrent response by the defender will be a relatively incredible indicator of resolve to fight, since an opportunistic defender can hope for "cheap deterrence" given the incredibility of the initial challenge.

This logic has an important consequence for empirical studies of deterrence. The

strategic behavior described above implies that *defenders' immediate deterrent threats will tend to be most credible indicators of intentions in cases where they are most likely to fail.* As a result of self-selection in crisis bargaining, immediate deterrent threats will tend to be relatively credible when the challenging state is more likely to be highly motivated, and thus ready to assume the risk of resistance by the defender. Conversely, immediate deterrent threats will tend to succeed more often in the cases in which they are relatively incredible—not because lack of credibility helps, but because in these cases the challenging states will tend to be opportunists looking for cheap concessions, and thus not willing to run much risk of a fight.[19] (Figure 3 outlines the major steps of the preceding argument.)

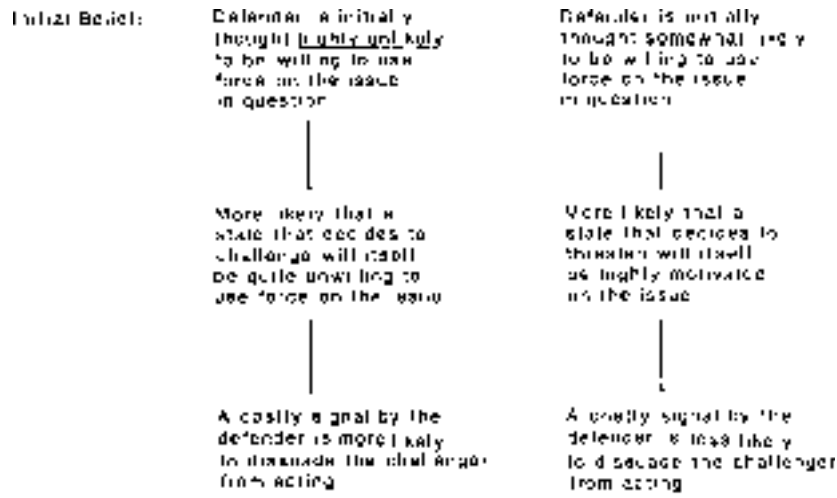

Figure 3. How challengers' prior beliefs influence the credibility and efficacy of immediate deterrent threats.

It follows that while classical rational deterrence theory may be right about general deterrence, it is wrong to apply it to immediate deterrence. Greater credibility may well make an initial challenge less likely. But to the extent that it does, then when a threat occurs this indicates a better chance of a strongly motivated challenger and a worse chance of immediate deterrence success.

Immediate deterrence is therefore not an empirically measurable, testable version of general deterrence. If costly signaling and the accompanying selection effects operate, then the cases of immediate deterrence that we observe will differ systematically from the cases of general deterrence that we do not. In fact, hypotheses that are valid for general deterrence should appear exactly reversed if we look at cases of immediate deterrence. For example, any factor that increases the chance of general deterrence success—such as an alliance or any other precrisis indicator of defender willingness to fight on behalf of a "protégé" state—should be associated with immediate deterrence failure in a sample of cases. Likewise, any precrisis indicator that

predicts a defender would not be willing to fight on behalf of a protégé should be associated with general deterrence failure but immediate deterrence success![20]

The argument does not say that greater credibility in a threat causes immediate deterrence failure, only that the two will be correlated across cases if states act strategically. If we could somehow hold all else equal in a particular historical case, the chance of immediate deterrence success should increase with the credibility of the defender's threat, in accord with the logic of the classical theory. Across cases, however, defender credibility will be associated with the challenging state's level of motivation and hence with immediate deterrence failure.

EMPIRICAL EVIDENCE AND ILLUSTRATIONS

According to the theoretical argument, if we take a set of international crises and then sort the cases into two groups according to the credibility of the defender's threats, we should find that immediate deterrence fails more often in the group in which the defender's counterthreat was more credible as an indicator of intent. This is essentially what Lebow and Stein do in their effort to "replicate [Huth and Russett's findings] through similar procedures." They discard a total of 43 cases from both the 1984 and 1988 data sets either for "lack of persuasive evidence of intention to attack" by the challenger, lack of a credible immediate deterrent effort by the defender, or both (Lebow and Stein, 1990a, p. 342). Consistent with the argument about selection effects, in the remaining cases immediate deterrence succeeds rarely (2 of 10 cases) and in no case for more than a few months.

While Lebow and Stein's findings are consistent with a major empirical implication of the theoretical argument, one may wonder about the extent to which the dynamics of costly signaling and selection effects are responsible. This section focuses on several cases from the data set, arguing that they show evidence of these dynamics at work.

The consequences of selection according to prior beliefs are perhaps easiest to see at the extremes—for instance, in cases where the challenging state initially thought resistance to be very likely prior to issuing its threat. Here the argument leads us to expect that an immediate deterrent threat by the defender will be highly credible but very unlikely to succeed.

The confrontation between Austria and Serbia in 1914 provides a nice example. Austrian leaders initially expected Serbia to reject their ultimatum – their *ex ante* belief was that Serbia would very probably be willing to use force rather than compromise its sovereignty (Albertini, 1953, pp. 164–178; Williamson 1991, p. 198). Given these initial beliefs, known by the Serbian government, the Austrian ultimatum credibly revealed an intention to act if not satisfied; the Serbs mobilized almost immediately in response. In turn, this Serbian counterthreat credibly indicated a willingness to resist with force. Yet despite the credibility of the response as an indicator of Serbian intentions, it was very unlikely to dissuade Austrian action since the Austrians had already taken it into account. Many similar cases can be adduced, a few of which are in the data set.[21]

At this extreme value for the challenger's initial beliefs, it does not seem particu-

larly surprising that the defending state's immediate deterrent threat is likely to fail even though it is a credible indicator of willingness to fight. What is less obvious is that this sort of case lies at the end of a continuum and that other cases on the continuum should reveal the same dynamic. The less the challenger's initial belief that the defender might be willing to resist with force, the less credible an immediate deterrent threat by the defender but the more likely it will be to work.

Comparisons within the case of July 1914 are useful for illustrating the theoretical argument across a range a initial beliefs. German (and Austrian) leaders held different prior beliefs about the likelihood that Serbia, Russia, and Britain, respectively, would be willing to go to war over the Austrian demands on Serbia. Whereas they expected that Serbia was almost certain to be willing to fight over the issue, they held a more moderate expectation about Russia's willingness, and seem to have had a still lower initial belief that Britain would be willing to intervene.

If rational self-selection into crises occurs, it will be important to distinguish such continuous variations in initial beliefs. The argument implies that a Russian immediate deterrent threat in response to the Austro-German challenge should have had a smaller chance of success than a British costly signal even though British–German signals would be less credible (i.e, less informative about actual intentions and more subject to discounting). In turn, Russian–German signals should have been relatively less informative and credible than Serbian–Austrian threats, but again more likely to work.

These predictions are consistent with broadly held views of the July crisis. In the first place, many have argued that war might have been averted if Lord Grey had sent an earlier, more forceful signal of British willingness to fight (Albertini, 1953, p. 514; Turner, 1970, p. 99). Soon after Grey finally did send the relatively costly signal received in Berlin on the night of the July 29, Chancellor Bethmann Hollweg cabled Vienna twice, urging restraint so that the worst-case war against Russia, France, and Britain might be avoided.[22] This effort was too late and too brief to have an effect – the Austrian leaders were too committed at this point for a quick reversal of course, and news of Russia's mobilization soon led Bethmann to abandon his effort to restrain Austria.[23]

At the same time, a number of scholars have noted that the Germans tended to discount the earlier signals they did receive, including several remarks by Grey between the 25th and 28th, and possibly also the British decision not to disperse the fleet at Scapa Flow as originally planned for July 28 (Lebow, 1981, pp. 131–132; Steiner, 1977, p. 226; Albertini, 1953, pp. 431–433).[24] Indeed, even Grey's relatively clear statement on July 29 did not lead the German leaders to completely revise their beliefs about Britain's willingness to fight. Right up to the British declaration of war on August 4, Germans leaders thought there was some chance England might not enter.[25]

The theoretical argument of the last section suggests that it may not have been unreasonable for the Germans to discount British signals to some degree. Given each side's initial beliefs—the Germans correctly thought the British Cabinet preoccupied with nascent civil war over Ireland, while Grey believed the Germans wished to avoid war as they had in past Balkan crises—an incentive existed to misrepresent one's willingness to use force on the issue. Indeed, when Lichnowsky began report-

ing that Grey's attitude suggested that Britain might actually fight with the Entente (July 26–27), Bethmann told his assistant of the "danger that France and England will commit their support to Russia in order not to alienate it, perhaps without really believing that for us mobilization means war, thinking of it as a bluff which they answer with a counterbluff" (Jarausch, 1969, p. 65).

By the selection effect argument, even if Grey had sent a more costly signal earlier on, Bethmann would have been right to remain skeptical about whether Britain would actually fight, due to Britain's incentive to misrepresent. Nonetheless, given Bethmann's reluctance to fight against the Entente plus Britain and the fact that the costs of shifting to a more conciliatory line were lower before the Austrian and Russian mobilization orders, even a somewhat discounted British costly signal may have been enough to have persuaded Bethmann to ease off. That is, a more costly immediate deterrent threat by the British may have had a reasonable chance of succeeding even though the Germans leaders could rationally have doubted whether it credibly indicated a British willingness to fight.[26]

The case of German–Russian signals is quite different, and the differences can be traced in part to different prior beliefs. In the counsels of state that approved the Austrian threat to Serbia, both Austrian and German leaders explicitly recognized and accepted the possibility of Russian intervention. Indeed, it was in a large part because of the perceived danger from Russia that, after Ferdinand's assassination, the Austrian leadership sent Count Hoyos as a special emissary to Berlin to inquire about German support (Albertini, 1953, pp. 124–125, 133, 138–139). In the discussions that led to the famous "blank cheque," both Kaiser Wilhelm and Bethmann Hollweg recognized a risk of Russian intervention, although the Kaiser at least was publicly guessing Russia would not intervene (Albertini, 1953, 138–139; Williamson, 1991, pp. 195–196).

Soon after these meetings on July 5th and 6th, Bethmann discussed the matter privately with his assistant, Kurt Riezler, to whom he confided his beliefs about the likelihood of different outcomes. According to Riezler,[27] Bethmann saw three possible outcomes for his diplomatic gamble: (1) a localized Austro–Serbian war with favorable consequences for Austro–Hungarian prestige and German diplomacy; (2) a continental war against Russia and France that Germany stood a reasonable chance of winning; and (3) a world war against Russia, France, and Britain that might well be disastrous for Germany. Bethmann apparently saw localization (Russia backs off) and a continental war (Russia fights) as roughly equally likely, while the third outcome of world war (Russia and Britain fight) was judged possible but less likely.[28]

If reliable, these estimates imply that Bethmann thought the odds that Russia would intervene were at least fifty–fifty! And even if we put less weight on this one conversation, Riezler's diaries along with much other evidence indicate that both in July and in the preceding months Bethmann was highly pessimistic about Germany's diplomatic situation. In particular, he saw that Russia was increasingly unwilling to tolerate Austria-Hungary's effort to conduct a "policy of prestige" in the Balkans (Jarausch, 1969, p. 53; Williamson, 1991, pp. 196–197). In sum, Bethmann did not give Austria–Hungary the "blank cheque" to crush Serbia with the prior expectation that Russia was highly unlikely to be willing to resist militarily.[29]

At the same time, Russian costly signals—the premobilization measures that be-

gan to be reported to Berlin on July 26—appear to have affected the beliefs of the German leaders more rapidly than did British actions. As the reports came in Bethmann appears to have concluded that Russia would probably fight; he immediately turned his diplomatic efforts to putting the blame on Russia for provoking war (Turner, 1970, pp. 100–102). And while doubt lingered about Britain's true willingness to enter the fray up to the August 4 declaration of war, Russia's partial mobilization, reported in Berlin on July 30, left little doubt among the German leaders that Russia would fight (Turner, 1970, p. 109). Russian military preparations were perceived as relatively credible indicators of Russian willingness to go to war.

In sum, comparisons of how Austrian and German leaders interpreted Serbian, Russian, and British signals in the July crisis tend to support the theoretical expectations about prior beliefs and strategic dynamics outlined in the last section. The less the prior belief that a defender would actually fight, the greater the potential impact of its immediate deterrent threats, despite their lesser credibility as indicators of willingness to go to war.[30]

The consequences of self-selection can also be seen in comparisons between more distant cases in the data set. As an example, consider the contrast between Germany/Austria versus Serbia/Russia in 1914, and Germany versus Venezuela/United States 1903.

In the 1903 dispute, the Germans threatened and began to use force to persuade the Venezuelan government to pay its debts and to compensate German nationals who lost property in a recent civil war. Based on earlier inquiries made in Washington, German diplomats had expected that the United States would be highly unlikely to resist on Venezuela's behalf.[31] Contrary to expectations, as they pressed their demands (with gunships) the United States concentrated tremendous naval power in the Caribbean while Theodore Roosevelt and his diplomats issued veiled warnings intended to induce Germany to shift from military pressure to the bargaining table. These military and diplomatic threats were effective—they led to settlement by arbitration—even though they cannot be regarded as highly credible as indicators of a willingness by Roosevelt to go to war with Germany.[32]

In both 1903 and 1914, German leaders selected themselves into a crisis by choosing to threaten. In 1914 they chose an issue that they knew entailed a serious risk of Russian intervention and war, a risk much greater than the risk anticipated in 1903 regarding the United States. Given this difference in initial beliefs, the selection effect argument predicts three important differences between the cases. First, German leaders should have had a greater value for war against Russia in 1914 than they did for war with the United States in 1903, which the evidence suggests is quite plausible.[33] Second, U.S. immediate deterrent threats should have been relatively less credible indicators of willingness to go to war than were Russian immediate deterrent threats in 1914. This again appears to be the case—witness the historical dispute over whether the U.S. actions should be understood as proper, credible threats at all. Finally, even though less credible, U.S. threats in 1903 should have had a better chance of succeeding than did Russian threats in 1914, due to the different levels of motivation of the challenger who had "selected into" the crises to begin with. This is consistent with the outcomes in each case and seems consistent as well with what is known about German diplomatic thinking in the two instances.

Case evidence thus tends to support the broader regularity indicated by Lebow and Stein's recoding of the whole data set: Rational self-selection into crises by challengers implies that immediate deterrent threats are more likely to succeed in the cases where they are relatively incredible indicators of the defending state's willingness to go to war. Conversely, they are more likely to fail in cases where they are relatively credible.

CRISIS BARGAINING VERSUS RATIONAL DETERRENCE THEORY

Beyond suggesting a theoretical resolution to the empirical dispute on extended immediate deterrence, the argument about signaling and selection effects has some broader implications for understanding how deterrence works in international confrontations. Most importantly, it makes clear that rational deterrence theory is not by itself a theory of crisis bargaining and for this reason fits badly when inflicted on actual cases.

The classical theory suggests criteria for immediate deterrence success that obscure what is happening in many cases in the data set. Rather than making simple binary decisions between "attack" and "don't attack," states are making decisions about the use of military force, *conditional on what they expect they can obtain through crisis bargaining.* When this is the case, it becomes difficult to say whether immediate deterrence succeeded or failed in the sense of the classical theory, because the classical theory's categories are ambiguous.

To see how the ambiguity arises, consider a case of extended immediate deterrence between a challenger, a defender, and a small protégé state. Suppose that following the challenger's initial threat, the defender threatens the challenger with war should the challenger attack. However, at the same time or soon after, the defending state forces the protégé to cut a deal with the challenger, perhaps involving territorial concessions. Suppose the challenger accepts the offer and does not attack. More broadly, imagine any international dispute in which a defender counters a challenge with both a deterrent threat and (perhaps at a later time) a set of concessions that the challenger accepts.

Does immediate deterrence succeed or fail in such cases? Classical rational deterrence theory cannot say. On the one hand, the challenger might have attacked if the defender had not made a credible immediate deterrent threat—thus a "success." On the other hand, the defender made substantial concessions under the threat of force, which the defender's effort at immediate deterrence did not avert—thus a "failure." The problem is that international disputes typically involve bargaining over a range of possible issue resolutions, and not just the attack/not attack and resist/not resist decisions considered by the classical theory. How large must a state's concessions be to render prior resistance a "deterrence failure?" How small must they be to constitute an "immediate deterrence success?" Classical deterrence theory cannot answer because it does not conceptualize the interplay between military threats and bargaining over a range of outcomes.

This theoretical problem plays an important although hidden part in the empirical dispute on extended immediate deterrence. It bears on the question of how one judges whether an immediate deterrent threat succeeded or was simply irrelevant. Under a

strict interpretation of the classical theory, an immediate deterrent threat succeeds if, in the absence of the threat, the challenger would have attacked. In other words, to justify the inference that immediate deterrence succeeded in a particular historical case, a researcher needs to make a counterfactual argument about what would have happened if no deterrent threat had been issued.

Lebow and Stein wish to discard many of Huth and Russett's cases on precisely these grounds: They find insufficient evidence to support the counterfactual claim that the challenger would have attacked had there been no deterrent threat.[34] To give a few examples, they argue for designating cases as "not deterrence encounters" because "Germany never seriously considered attacking France [over Morocco in 1911];" "Serbian officials never contemplated attacking Austria [in a 1912 crisis];" "Japan had no immediate plans to invade Mongolia [in 1936];" "[Mussolini] did not seriously contemplate or prepare to attack France at this time [1938-1939];" "there is no evidence the Soviets were planning an attack [on Turkey in 1947];" "considerable uncertainty surrounds [Iraqi] premier Qasim's intentions. . . . [His threat to attack Kuwait in 1961] was probably a bluff."[35]

But if states have more options in international disputes than simply attack/don't attact and resist/don't resist, then it is not clear that the use of force by the challenger is the only counterfactual event that can support an inference of "immediate deterrence success." Military attacks are typically risky and costly. Even if the challenger would be willing to use force under some circumstances, why should it choose a costly military attack if it might have what it wants given to it in the form of tacit or explicit concessions? For example, without the defender's immediate deterrent threat, the protégé might have made significant concessions in bargaining rather than face war with the challenger, concessions that even a "serious" challenger might have preferred to a costly attack. When such a counterfactual scenario is plausible—coerced concessions by the protégé or the defender rather than simply attack by the challenger—it does not seem unreasonable to argue that immediate deterrence "succeeded." In such cases a defender's costly signal of willingness to resist may have resulted in significantly fewer concessions than would have been made if the defender had not tried immediate deterrence.

At times it seems that Lebow and Stein would like to rule such cases out of the data set. When criticizing Huth and Russett's coding they assume that immediate deterrence succeeds only when the challenger would have attacked immediately had there been no threat by the defender. However, in their own formal definition of immediate deterrence success they seem to acknowledge the issue raised above and the coding ambiguity it implies. They hold that immediate deterrence succeeds if the "challenger considered an attack *or a proscribed action,* but decided against proceeding because the defender persuaded the challenger that there would be serious and unacceptable consequences" (Lebow and Stein, 1990a, p. 345, emph. added). "Proscribed actions" are a much broader class than military attacks, and would seem to cover coerced or tacit concessions to the challenger from protégé or defender. Proscribed actions of this sort are certainly consistent with Huth and Russett's coding rules—they define immediate deterrence failures to include not only attacks but also cases where the defender made substantial concessions to the challenger "under the threat of force," as in the 1938 Munich crisis. By this definition, successful im-

mediate deterrence need not ward off an imminent attack, but only some further pressure, action, or a concession by the defender (Huth, 1988, p. 27).[36]

In sum, classical rational deterrence theory appears to suggest relatively clear criteria for coding deterrence failure, thus making possible clear empirical tests. However, when one actually tries to apply the criteria to empirical cases, one encounters the problem discussed above—the use of military threats in crises is bound up with bargaining over a range of possible issue resolutions, and this fact renders determination of deterrence "success" and "failure" problematic under the classical theory's categories. If an immediate deterrent threat averted significant coerced concessions to the challenger, then it seems reasonable to say that immediate deterrence succeeded. But how large must the counterfactual concessions have been—or, what is almost the same thing, how strongly motivated the challenger coercing them—for the case to qualify as a "deterrence encounter"? Lebow and Stein implicitly set the threshold high, sometimes allowing only for immediate attack. Huth and Russett implicitly set a lower threshold.[37] Neither decision can be justified by classical rational deterrence theory, whose categories simply do not fit into actual cases of crisis bargaining. Once again, what looks to be an empirical dispute about historical interpretation and coding actually derives from problems with the theory being tested.

The costly signaling framework sketched above is somewhat better equipped to conceptualize the interplay between military threats and bargaining in international disputes.[38] Why do states use military threats at all? Since the use of force is typically risky and costly, states in dispute have strong incentives to learn whether there are agreements both sides would prefer to fight. But because leaders have private information about what they are willing to fight over, and because they can have incentives to misrepresent this information in bargaining, quiet diplomatic conversations may be insufficient to allow learning about what another state is or is not willing to concede. Instead, costly signals are required. Under certain conditions military threats fit this bill.

An international crisis may be thought of as a sequence of costly signals issued by challenger and defender. With each signal by the opponent each state should grow more confident that the opponent will not make concessions, or should decrease its estimate of the maximum concessions the opponent will make. This is because with each round of escalation, a state with low resolve on the issues is more likely to make concessions than a high resolve state.

Military threats by a defending state in a crisis thus do more than simply attempt to deter potential attack. Rather, *they communicate (noisily) what deals the defender will or will not accept in preference to war.* A rational challenger should not even consider attack until it has learned from prior threats and signaling that it is unlikely to obtain what it wants by diplomatic concessions.[39] If a crisis evolves to the point where the challenger is seriously considering attack—say, drawing up war plans—the challenger has probably already learned that sufficient coerced concessions are relatively unlikely. As argued earlier, a challenger with such beliefs who has reached this point is unlikely to be dissuaded by immediate deterrent threats that follow, even if they are relatively credible. By contrast, if the challenger chooses to stop pressing its demands before it reaches the point of drawing up war plans, this does not mean that the defender's immediate deterrent threats were irrelevant—the outcome had

immediate deterrence not been tried may have been concessions extracted under threat of force by the challenger.

CONCLUSION

Specifying and testing deterrence theory are not two separate enterprises. Theory influences how empirical evidence is assembled, classified, and interpreted. In the debate over immediate deterrence, both sides rely on classical rational deterrence theory to structure data on international crises and to guide its interpretation. I have argued, however, that while the classical theory may be relevant when applied to general deterrence it is inappropriate for the analysis of immediate deterrence due to the strategic implications of self-selection by challengers and defenders into crises. Without a theoretical understanding of these strategic implications, what the data "says" about deterrence is likely to be misunderstood. I conclude with two points about how the data on immediate deterrence should be interpreted in light of selection effects and a more general point about the methodological signficance of selection effects.

First, contrary to their suggestion that the data are inconsistent with rational behavior by states, Lebow and Stein's reading of the evidence actually lends some support to a rationalist theory of crisis bargaining. If crisis signals are costly, then rational self-selection implies that across cases, the credibility of the defender's immediate deterrent threats will be correlated with the challenging state's motivation to overturn the status quo, and hence with immediate deterrence failure. Because Lebow and Stein code Huth and Russett's data for the credibility of challenger and defender threats as indicators of intentions, they find that immediate deterrence almost never succeeds.

Second, it is a mistake to conclude from the data, as Lebow and Stein do, that immediate deterrent threats almost never have their intended effect. It may be true that immediate deterrence will rarely prevent attack by a state whose leaders have concluded, through prior negotiation or crisis bargaining, that sufficient concessions are unlikely. But if in most disputes states bargain over a range of possible resolutions, then immediate deterrent threats may play an important role in revealing what negotiated outcomes a state might reject in preference for force. The alternative to risking immediate deterrence may be acquiescing in concessions that the challenger can obtain at lower cost than by military attack. This simple point is easily missed if one sees crises through the lens of classical rational deterrence theory, which focuses on the decision to attack and neglects bargaining over a range of issue resolutions.

Finally, while it is tempting to understand the impact of selection effects in crisis bargaining as selection bias—a statistical problem that might be remedied—doing so makes sense only as long as we are committed to using evidence on immediate deterrence to draw inferences about "deterrence in general." If cases of immediate deterrence represent a sample that is biased due to selection effects, then what is the relevant population from which it is drawn? The classical theory led us to imagine a relevant class of "all deterrence encounters" including both general and immediate (e.g., Huth, 1988, p. 17). I have argued that it makes more sense to think about general and immediate deterrence as two decision points within a single model of crisis

bargaining. From this vantage, the two decisions are so integrally linked that it is misleading to view one as just a biased version of the other. Instead of trying to control for the effects of "selection bias," studies of deterrence might better develop hypotheses that take account of selection effects and test directly for their impact.[40]

This point may apply more broadly. There are good reasons to expect that selection effects are common, important, and dangerous to neglect in studies of comparative and international politics. Often the cases we observe—of immediate deterrence, war, revolution, democratic transition or breakdown, and so on—are the result of choices made by people who would have chosen differently if circumstances had been different. And often the factors that influence their choices (such as prior beliefs about a defender's resolve) also influence the outcome in question. Thus the cases we observe rarely arise from anything like the process of random assignment of independent variables assumed by statistical models or, implicitly, in historical case studies. If we do not understand how the selection process that generates cases works, we are likely to draw wrong inferences about the cases and the theories under evaluation. A natural way of avoiding this is to theorize explicitly and to trace empirically the choices that make cases of nothing into cases of something.

NOTES

1. For the debate, see Lebow and Stein (1990a) and Huth and Russett (1990). Lebow and Stein assessed only 51 of the cases Huth and Russett included in their final version of the data set because they began work on an earlier version. For the data sets, see Huth and Russett (1984), Huth and Russett (1988), and Huth (1988).
2. While concurring on the core definition of immediate deterrence, Huth and Russett (1990, p. 468) argue that in applying it to the cases Lebow and Stein are led astray by "conceptual imprecision and theoretical misunderstandings." These concern primarily the distinction between deterrence and compellence and the role of uncertainty in deterrence.
3. The distinction between general and immediate deterrence was proposed by Patrick Morgan (1977).
4. On selection bias, a statistical problem that can result from selection effects, see, for example, Achen (1986), Heckman (1990), Przeworski, Alvarez, Cheibub, and Limongi (2001, App. 1). Regarding the study of deterrence, Achen and Snidal (1989, pp. 160–61), Levy (1989, p. 118), Huth and Russett (1990, p. 480n), Huth and Russett (1993, p. 62), and Lebow and Stein (1990b, pp. 9–10) mention selection bias as a problem for attempts to estimate the rate of general deterrence success from a sample of failures. I argue here that the strategic logic behind selection effects in the case of immediate deterrence has not been understood and the full empirical implications not appreciated.
5. For contrary views, see Telhami (1992) and Stein (1992).
6. Lebow and Stein discard several cases on the ground that there is insufficient evidence available to code them.
7. Three plus eight is eleven, not ten: Lebow and Stein originally saw two cases of extended immediate deterrence in the Munich crisis, where Huth and Russett saw one (Lebow and Stein 1990a, pp. 363–365). In their final revision, Lebow and Stein decided to code Munich as a single instance of failure, for reasons they do not elaborate. This must reflect some sort of overall judgement on the case rather than a new historical interpretation, since their account of how the crisis played out is unchanged. See Lebow and Stein (n.d., case #25). Henceforth I will refer to cases in this set of case summaries as LS#x, where x is the case number. Likewise, Huth and Russett's case summaries will be referred to as HR#x (Huth n.d.).
8. At the level of rational deterrence theory what matters is the expected cost of the sanction, not whether the threat is military or economic. Huth and Russett justifiably limited their sample to military threats to save effort.
9. Huth and Russett argue in several places that proper coding of challenger and defender intentions is

critical. For example, their coding procedure involved "searching for evidence of alternative reasons why the threats were issued purely as bluffs" (p. 483). In Huth and Russett (1988, p. 31) they state that "To distinguish between bluffs and true intentions to attack we consulted closely with diplomatic historians, country experts, and secondary analyses to ascertain the best available judgement of the 'attacker's' apparent intentions and goals." However, Huth and Russett (1990, p. 479) have also argued that cases of "bluff and probing" should be included in a valid sample, and that Lebow and Stein misunderstand deterrence theory when they use seriousness of intentions as a coding criterion. Regardless of what one makes of this ambiguity, it is fairly clear that the two sides get different results largely because of the different approaches they take on the issue of intentions.

10. In the international relations literature, the cost to the challenger of resistance by the defender, or the military efficacy of that resistance, is often misleadingly included in the concept of "credibility" as well. Here the word will refer to the likelihood that the threat would actually be carried out, and not to how painful or effective the proposed punishment would be.
11. To illustrate, the July crisis of 1914 is coded as follows: (1) the Austrians and Germans decide to threaten Serbia rather than accept the status quo; (2) the Russians mobilize and issue counterthreats on Serbia's behalf; (3) the Austrians and Germans choose to attack Serbia and to initiate an attack on Russia (and her ally, France); (4) Russia chooses to use force in response. For Huth and Russett's coding rules, see Huth (1988, p. 23–27).
12. Informal versions of the model in Figure 2 have appeared in other empirical studies of crises as well, such as George and Smoke (1974, pp. 101–103).
13. Though related, costly signaling differs from the notion of commitment developed by Schelling (1960). Costly signaling is a means for revealing unobservable preferences when there are incentives to misrepresent them; commitment tactics are a means for observably rearranging the incentives one will face in a future contingency, when one's preferences are known or almost certain. See Fearon (1992, ch. 3) for a discussion. For the original analysis of costly signaling in economics, see Spence (1973).
14. This condition is not always met. For example, one reason that it was difficult to discern whether Saddam Hussein was actually willing to fight from his pronouncements in the Fall of 1990 was that it was hard to say whether he would suffer serious audience costs for bluffing. More generally, when a small state confronts a much bigger one, the leaders of the small state sometimes may actually be applauded by domestic audiences for "standing up to the bully," even if they ultimately back down. Note also that relevant domestic audiences may be as large as a mass public or as small as a politburo—while audience effects are probably greater on average in democracies, they frequently operate in non-democracies as well (Fearon, 1994a).
15. Morrow (1989) was the first to observe selection effects in a game–theoretic model of crisis bargaining, though his model considers states' offers and counteroffers rather than the military threats and signals involved in deterrence. See also Banks (1990).
16. This is not the case with a costless signal—also known as "cheap talk"—that provides an opportunistic state no disincentive to taking the same actions as a motivated state would, so that signal conveys no information about true willingness to fight.
17. Throughout, I use "motivation" and "resolve" interchangeably in order to facilitate making the connection between the game–theoretic argument and the debate on extended immediate deterrence. I take resolve to be a state's willingness to use force over specific interests—in formal terms, a state's expected utility for military conflict as against its utility for possible negotiated settlements, including the status quo. Thus "resolve" incorporates (a) the military balance (the probability of winning a fight); (b) the state's utility for winning or losing on the particular issues at stake; and (c) the state's expected costs for a fight. In the literature, "motivation" sometimes refers to all three, sometimes to (b) and/or (c), and sometimes to a more general preference concerning conquest and expansion.
18. In the literature, the impact of prior beliefs on how leaders interpret signals has been analyzed almost entirely in terms of psychological biases (Jervis, 1976; Lebow, 1981).
19. One might wonder if the greater credibility of the defender's threat would simply offset the challenger's higher motivation, implying no relation across cases between credibility and the likelihood of success. But the greater credibility of the defender's immediate deterrent threat is anticipated by the challenger prior to threatening, so it has no added impact when made.
20. Huth's several indicators of the strength of the defender's interest in the protégé provide some material for testing these hypotheses. Using data generously provided by Paul Huth, I found that two

indicators that predict the defender would fight on behalf of the protégé are associated with immediate deterrence failure (alliance and geographic proximity). Surprisingly, military transfers between defender and protégé predict that the defender will not fight and, as anticipated by the selection effect argument, they also predict immediate deterrence success. Finally, the level of trade between defender and protégé acts anomalously: Higher levels of trade predict that the defender will fight and also immediate deterrence success. See Fearon (1994b). In their 1984 article, Huth and Russett had found that alliances are associated with immediate deterrence failure, and explained this with the core of a selection effect argument. Jervis (1991, pp. 121–22) has also noted the possibility. But the full logic and implications of the argument have not been seen. For example, if the argument applies to alliances between defender and protegé, it should also apply to any other indicator of defender interest that is available before an initial challenge. We cannot use the argument for alliances and then abandon it for trade ties or geographic proximity.

21. For example, Italy vs. Tripoli/Turkey 1911 (LS #9, HR #14); Bulgaria vs. Greece/Serbia 1913 (LS #15, HR #19); India vs. Goa/Portugal 1961 (LS #39, HR #44).
22. The news from Grey was reported in a telegram from Ambassador Lichnowsky (Kautsky 1924, No. 368). Among other things, Grey had told Lichnowsky that if France became involved in a war, "it would be impracticable to stand aside and wait for any length of time." Given the constraints Cabinet government implied for Grey's control of foreign policy, this was a costly signal (Fearon, 1992, pp. 147–151). Bethmann's cables to Vienna are given in Kautsky (1924, Nos. 395–396). These were dispatched together with a reply to Grey, thanking him for his "frank explanation" of the British position and noting that he was "urgently advising" Vienna to accept Grey's mediation proposals (No. 393). Trachtenberg (1991, pp. 85–87) has argued that news of the Russian partial mobilization more than the British position inspired Bethmann's brief change of course. I agree with Levy (1990/91, pp. 165–166) that the evidence tends to support the more conventional view here.
23. In an unsent version of a July 30th telegram to his ambassador in Vienna, Bethmann wrote "I canceled the order of instructions [asking Vienna to accept mediation] as the General Staff just informs me that the military preparations of our neighbors, especially in the east, will force us to a speedy decision, unless we do not wish to expose ourselves to the danger of surprise" (Kautsky, 1924, No. 451). (A shorter version with the same instructions was sent.)
24. Since Grey was actively pushing mediation proposals and clearly avoiding a strong statement in this period, it cannot be said that a particularly costly signal was sent at this time. The instructions on the fleet, published in London on the 27th, would count on this score, but it is not clear when or what Berlin heard about this decision (Albertini 1953, p. 429). Even so, Grey's failure to more clearly indicate neutrality—his "staying in" rather than "dropping out"—aid something about British intentions. The signals preceding the July 29th statement appear to have somewhat increased the German leaders' beliefs that England might fight, as evidenced by Bethmann's increasing diplomatic efforts to make Russia seem the aggressor and to persuade Britain to stay out (Albertini, 1953, 443ff.). See also Jarausch (1969, p. 65) on Riezler's diary for these days.
25. On August 2, for example, Bethmann asked Lichnowsky to relay evidence of French aggression against Germany and to convince Lord Grey that "Germany, although she has been advocating the maintenance of peace to the utmost limit, is being driven by her opponents to adopt the role of an injured party who must take to arms, for the preservation of her very existence" (Kautsky, 1924, No. 693; Albertini, 1953, vol. 3, pp. 50–51). Late on August 3, Lichnowsky cabled Berlin that "the local Government has no immediate intention of departing from its neutral stand," and that "I am convinced that the local Government will strive to remain neutral" (Kautsky, 1924, No. 801; see also Nos. 742, 744, and 810).
26. I should stress that a substantial part of German uncertainty about British intentions was due not solely to British private information but also to environmental uncertainty—uncertainty arising from mutual ignorance about the outcome of effectively random processes (Fearon, 1992, ch. 3). Grey himself emphasized on several occasions that British public opinion would be decisive; like the Germans, the British leaders were quite uncertain about how this factor would turn out. As late as August 1 Grey was telling the horrified French Ambassador of Cabinet deliberations implying that "France must take her own decision at this moment without reckoning on an assistance that we were not now in a position to promise" (Albertini, 1953, vol. 3, p. 392). At the same time, Grey was making remarkable offers of neutrality to Lichnowsky conditional on Germany not attacking France (Albertini 1953, vol.

3, pp. 380–386). These offers, which were not very well thought out, were probably in part an attempt by Grey to find some way to out of the Cabinet crisis then taking place over the possibility of war.

27. More precisely, this is historian Konrad Jarausch's recapitulation of Riezler's diary (Jarausch, 1969, esp. pp. 58–59).
28. Bethmann explicitly told Riezler that "An attack on Serbia can lead to world war" (p. 58).
29. On July 20 Bethmann told the Kaiser that "the solution of this problem [of localizing the conflict] is so difficult that even a minor incident can tip the scales." On the 23rd he told Riezler that "Should war break out, it will result from Russian mobilization *ab irato,* before possible negotiations. In that case we could hardly sit and talk any longer, because we have to strike immediately in order to have any chance of winning at all" (Jarausch, 1969, pp. 62–63).
30. For at least two reasons, I should emphasize that these results are supportive rather than decisive. First, the claims involve some counterfactual argument (what would have been the effect of an earlier costly signal by the British?). Second, the partial interdependence of Serbian, Russian, and British choices complicated Germany's decision problem in ways not addressed by the simple model of Figure 2; for example, the Germans could reasonably think that the Russian decision to fight might depend in part on the British decision.
31. In November 1902, U.S. Secretary of State Hay told the British and the Germans that while the United States regretted the use of force, this would not be opposed as long as no territorial acquisitions followed (Healy, 1988, p. 101; Collin, 1990, p. 93).
32. Collin (1990, ch. 4) is good recent account. As noted in the Lebow/Stein and Huth/Russett summaries of this case, there is some dispute over whether Roosevelt issued an ultimatum in private to the German ambassador. Whether he did is immaterial to the argument made here.
33. See Collin (1990) and sources on World War I cited above. Recall that by the selection effect argument, the greater the initial expectation of resistance, the greater the level of motivation required to make a challenge rational.
34. Oddly, Lebow and Stein (1990a, p. 343n7) disavow counterfactual argument, even though this approach underlies their critique of Huth and Russett.
35. These statements refer to the following cases: Germany v. Morocco/France 1911, LS #9; Serbia vs. Austria/Germany 1912, LS #13; Japan vs. Mongolia/Soviet Union 1936, LS #24; Italy v. Tunisia/France 1938-9, LS #26; Soviet Union vs. Turkey/U.S. 1947, LS #31; Iraq vs. Kuwait/Britain 1961, LS #38.
36. The researchers' dispute over coding Munich 1938 illustrates: The case was originally coded as an immediate deterrence failure by Huth and Russett but as a partial success by Lebow and Stein. Lebow and Stein correctly note that Hitler *did* want to attack Czechoslovakia and seems to have been dissuaded in part by the prospect of war signaled haltingly by Britain and France. Huth and Russett correctly note that despite commitments made in May on behalf of Czechoslovakia, in September Britain and France made major concessions to Hitler due to his threats to use force. If the relevant counterfactual event is an attack ordered by Hitler, then immediate deterrence succeeded. If the criterion admits "proscribed actions" involving demands on Czechoslovakia and the occupation of the Sudetenland, then immediate deterrence failed. In fact, an imminent attack was simultaneously deterred by threats and bought off by concessions. See HR #31; LS #25; and Taylor (1979, pp. 896–897).
37. For Huth and Russett, the problem posed by the interplay of threats and bargaining expresses itself in ambivalence over how to treat bluffing by challengers; see footnote 9 above.
38. However, the model of Figure 2 does not formally characterize a bargaining process in which states can make continuous offers and counteroffers. In Morrow's (1989) model, states can choose one of two offers in addition to conceding, although the relationship between offers and military threats in the game is unclear. Fearon (1995) considers a take-it-or-leave-it bargaining game with one state making a continuous offer/demand and the other either accepting or using force in reply. Powell (1996) analyzes an infinitely repeated version of this game, and finds that, surprisingly, its (essentially) unique equilibrium involves one state making a take-it-or-leave-it offer.
39. Assuming that fighting is not valued for its own sake or for domestic effect by the state's leaders.
40. I attempted to do this with Huth and Russett's data in Fearon (1994b).

REFERENCES

Achen, Christopher (1986). *Statistical Analysis of Quasi-Experiments,* Berkeley: University of California Press.

Achen, Christopher and Duncan Snidal (1989). "Rational Deterrence Theory and Comparative Case Studies." *World Politics,* Vol. 41, pp. 143–170.

Albertini, Luigi (1953). The Origins of the War of 1914, Vol. 2. London: Oxford University Press.

Banks, Jeffrey S. (1990). "Equilibrium Behavior in Crisis Bargaining Games." *American Journal of Political Science,* Vol. 34, pp. 599–614.

Collin, R. H. (1990). *Theodore Roosevelt's Caribbean,* Baton Rouge: Louisiana State University Press.

Ellsberg, Daniel (1975). "The Theory and Practice of Blackmail." In *Bargaining: Formal Theories of Negotiation*, ed. Oran Young, Urbana: University of Illinois Press, pp. 343–363.

Fearon, James D. (1990). "Deterrence and the Spiral Model: The Role of Costly Signals in Crisis Bargaining." Paper presented at the Annual Meetings of the American Political Science Association, San Francisco, 31 August.

Fearon, James D. (1992). Threats to Use Force: Costly Signals and Bargaining in International Crises. PhD thesis, U. of California, Berkeley.

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

Fearon, James D. (1994b). "Signaling versus the Balance of Power and Interests: An Empirical Test of a Crisis Bargaining Model." *Journal of Conflict Resolution,* Vol. 38, pp. 236–69.

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

George, Alexander and Richard Smoke (1974). *Deterrence in American Foreign Policy,* New York: Columbia University Press.

Healy, D. (1988). *Drive to Hegemony: The U.S. in the Caribbean, 1898–1917,* Madison: University of Wisconsin Press.

Heckman, James (1990). "Selection Bias and Self-Selection." In *The New Palgrave*: Econometrics, ed. John Eatwell et al., New York: Norton.

Huth, Paul (1988). *Extended Deterrence and the Prevention of War,* New Haven: Yale University Press.

Huth, Paul (n.d.) "Appendix (Case Summaries)." Typescript, University of Michigan.

Huth, Paul and Bruce Russett (1984). "What Makes Deterrence Work? Cases from 1900 to 1980." *World Politics,* Vol. 36, pp. 496–526.

Huth, Paul and Bruce Russett (1988). "Deterrence Failure and Crisis Escalation." *International Studies Quarterly,* Vol. 32, pp. 29–46.

Huth, Paul and Bruce Russett (1990). "Testing Deterrence Theory: Rigor Makes a Difference." *World Politics,* Vol. 42, pp. 466–501.

Huth, Paul and Bruce Russett (1993). "General Deterrence Between Enduring Rivals." *American Political Science Review,* Vol. 87, pp. 61–74.

Jarausch, Konrad (1969). "The Illusion of Limited War: Bethmann Holweg's Calculated Risk, July 1914." *Central European History,* Vol. 2, pp. 48–76.

Jervis, Robert (1976). *Perception and Misperception in International Politics,* Princeton: Princeton University Press.

Jervis, Robert (1991). "Systems Effects." In *Strategy and Choice,* ed. Richard Zeckhauser, Cambridge: MIT Press.

Kaufmann, William (1954). *The Requirements of Deterrence*, Princeton: Center of International Studies.

Kautsky, Karl, ed. (1924). *German Documents Relating to the Outbreak of the World War,* New York: Oxford University Press.
Lebow, Richard Ned (1981). *Between Peace and War,* Baltimore: Johns Hopkins University Press.
Lebow, Richard Ned and Janice Gross Stein (1987). "Beyond Deterrence." *Journal of Social Issues,* Vol. 43, pp. 5–72.
Lebow, Richard Ned and Janice Gross Stein (1989). "Rational Deterrence Theory: I Think, Therefore I Deter." *World Politics,* Vol. 42, pp. 208–224.
Lebow, Richard Ned and Janice Gross Stein (1990a). "Deterrence: The Elusive Dependent Variable." *World Politics,* Vol. 42, pp. 336–369.
Lebow, Richard Ned and Janice Gross Stein (n.d.) "Review of Data Collections on Extended Deterrence by Paul Huth and Bruce Russett." Unpublished typescript.
Lebow, Richard Ned and Janice Gross Stein (1990b). *When Does Deterrence Succeed and How Do We Know?,* Ottawa: Canadian Institute for International Peace and Security.
Levy, Jack S. (1989). "Quantitative Studies of Deterrence Success and Failure." In *Perspectives on Deterrence,* ed. Paul Stern et al., New York: Oxford University Press.
Levy, Jack S. (1990/91). "Preferences, Constraints, and Choices in July 1914." *International Security,* Vol. 15, pp. 151–86.
Morgan, Patrick (1977). *Deterrence: A Conceptual Analysis,* Beverly Hills: Sage Publications.
Morrow, James D. (1989). "Capabilities, Uncertainty, and Resolve: A Limited Information Model of Crisis Bargaining." *American Journal of Political Science,* Vol. 32, pp. 941–972.
Powell, Robert (1996). "Bargaining in the Shadow of Power." *Games and Economic Behavior,* Vol. 15, pp. 255–289.
Przeworski, Adam, Michael E. Alvarez, Jose A. Cheibub, and Fernando Limongi (2001). *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990,* Cambridge: Cambridge University Press.
Schelling, Thomas C. (1960). *The Strategy of Conflict,* New Haven: Yale University Press.
Snyder, Glenn (1961). *Deterrence and Defense,* Princeton: Princeton University Press.
Spence, A. Michael (1973). "Job Market Signalling." *Quarterly Journal of Economics,* Vol. 87, pp. 355–374.
Stein, Janice Gross (1985). "Calculation, Miscalculation, and Conventional Deterrence." In *Psychology and Deterrence,* ed. Robert Jervis et al., Baltimore: Johns Hopkins University Press, pp. 34–88.
Stein, Janice Gross (1992). "Deterrence and Compellence in the Gulf, 1990–91: A Failed or Impossible Task." *International Security,* Vol. 17, pp. 147–179.
Steiner, Zara (1977). *Britain and the Origins of the First World War,* London: MacMillan.
Taylor, Telford (1979). *Munich: The Price of Peace,* Garden City, NJ: Doubleday.
Telhami, Shibley (1992). "Between Theory and Fact: American Behavior in the Gulf War." *Security Studies.*
Trachtenberg, Marc (1991). *History and Strategy,* Princeton: Princeton University Press.
Turner, L.C.F. (1970). *Origins of the First World War*, New York: Norton.
Williamson, Samuel (1991). *Austria–Hungary and the Origins of the First World War,* New York: St. Martin's.

CONTRIBUTOR

James D. Fearon is Professor of Political Science at Stanford University. His research has focused on democracy and interational disputes, explanations for interstate wars, and, most recently, the causes of civil and especially ethnic violence. He is presently working on a book manuscript (with David Laitin) on civil war since 1945.