Challenges to Inference in the Study of Crisis Bargaining*

Philip Arena
University at Buffalo, SUNY
Department of Political Science
parena@buffalo.edu

Kyle A. Joyce
University of California, Davis
Department of Political Science
kjoyce@ucdavis.edu

October 19, 2011

*We thank Amber Boydstun, Daina Chiba, Sean Gailmard, Heather McKibben, Jamie Monogan III, Bill Reed, Jas Sekhon, and Curt Signorino for valuable comments. We thank Matt Buttice for research assistance. This article has also benefited from comments provided by the participants at EITM (Berkeley, 2010) and attendees of the annual meeting of the Society for Political Methodology (2010).

ABSTRACT

The possibility that actors strategically condition their behavior on partially unobservable factors poses an even graver challenge to causal inference than has been previously appreciated. In order to examine the challenges, we present a simple crisis bargaining model that indicates that targets can generally prevent war by arming. We then create a simulated data set where the model is assumed to perfectly describe the onset of war for those states engaged in crisis bargaining, which we assume most pairs of states are not. We further assume researchers cannot observe which states are engaged in crisis bargaining, although observable variables might serve as proxies. We demonstrate that a naïve design would falsely indicate a positive relationship between arming and war. We then evaluate the ability of matching, instrumental variables, and statistical backwards induction to uncover the true negative relationship. We show that each method faces limitations, which are generally worse for less theoretically motivated approaches.

1. INTRODUCTION

It is well understood that if some units face fundamentally different risks of exposure to a given outcome, then failure to control for the factors that account for such differences can bias our estimates of the effects of other variables. If the factors that account for different risks of exposure are observable, the remedy is straightforward. We might control for those factors, pre-process our data with matching methods, or try to identify instruments for our potentially endogenous regressors. But what if the factors that account for different risks of exposure are unobservable? In this case, we can turn to proxy variables (i.e., observable factors that we have reason to believe will correlate with those that are believed to be generating omitted variable bias). However, there is no guarantee that we will draw the correct inference if we are only able to partially account for the bias.

There are two reasons that scholars of international conflict in particular should find this prospect quite troubling. First, the most prominent explanation for war implies that some pairs of states will indeed face systematically different risks of exposure to international conflict, and that the source of this variation is largely unobservable. Second, we do not know much about how advanced statistical techniques that are typically employed to address such problems perform under the conditions that applied researchers likely face in practice.

With respect to the first point, a growing number of scholars view war either as the result of bargaining failure[1] or an extension of the bargaining process.[2] Regardless of whether conflict occurs after bargaining ends or as part of the bargaining process itself, such accounts stand in stark contrast to the "all-or-nothing" nature of traditional approaches to international conflict.[3] Most notably, when issues are assumed to be indivisible, we typically assume that crises begin when one state makes a publicly observable challenge to the status quo, typically envisioned as a public threat or a militarized act.

[1] See especially Fearon (1995), Powell (1999) and Tarar and Leventoğlu (2008).
[2] See Wagner (2000), Slantchev (2003), Powell (2004) and Filson and Werner (2002).
[3] See, for example, Zagare and Kilgour (2000), Slantchev (2011), and Signorino and Tarar (2006).

This view appears to implicitly inform the major data sets analyzed by scholars of international conflict. For example, the Militarized Interstate Dispute (MID) data set records all incidents in which states engage in militarized activity towards one another, whether it is threats to use force, shows of force, or actual fighting (Ghosn, Palmer, and Bremer 2004).[4] Many scholars use this data set to test the implications of theoretical models in which goods are assumed to be indivisible. In so doing, they assume that MID initiation represents a challenge to the status quo, and MID reciprocation indicates crisis escalation (Sartori 2005, Schultz 2001).

The bargaining models of war challenge the assumption that wars occur only after a chain of publicly observable hostile actions were taken. Such models typically assume that states issue ultimata to one another, and that these may be delivered in private or simply be implicit in their actions, as in the case of a *fait accompli*. Moreover, such models typically identify what is known as the risk-return tradeoff as the cause of war. Challengers are assumed to be unable to determine whether the target will accept or reject any given terms, so their proposal must balance the risk of war against the benefit of securing a better agreement. Strictly unobservable factors will therefore often separate 1) peaceful incidents in which a state unilaterally altered the status quo in some way, without having engaged in a threat, display, or use of force, yet *ex ante* understood that their actions created a positive risk of war, from 2) cases in which wars do in fact occur.[5]

Put simply, one of the most prominent approaches to explaining war suggests that we cannot assume that "crises" always begin with observable actions of the sort recorded by existing data sets. For all we know, a great many near-misses go unrecorded. We have no way of knowing whether this is a rare occurrence or a common one, but the mere possibility implies that existing data sets provide an imperfect basis for determining which dyads (pairs

[4]The International Crisis Behavior (ICB) data set, the other major data set commonly used by scholars of international conflict, records all incidents in which there is observable evidence that actors perceived 1) a threat to their basic values, 2) time pressure for response, and 3) a heightened probability of military hostilities (Hewitt 2003).

[5]See Gartzke (1999) for further discussion of this point and its implications for empirical inquiry.

of states) have the potential to come into conflict with one another.

As we demonstrate below, an inability to distinguish between dyads that are engaged in crisis bargaining from those that are not may pose a profound challenge to our ability to correctly diagnose the relationship between the actions states take and the likelihood that they subsequently find themselves at war. Any behavior states undertake when engaged in crisis bargaining in order to prevent escalation, but would rarely undertake otherwise, will be positively correlated with the incidence of conflict.

The typical solution to concerns about omitted variable bias, of course, is to include a laundry list of observable control variables, although there are problems with this approach (Clarke 2009). We demonstrate that even when there is a valid concern about omitted variable bias, the inclusion of proxy variables, which is often the only option we have available, may not ensure correct inferences.[6]

With respect to the second point above, we further demonstrate that even advanced statistical techniques such as matching, instrumental variables, and structural estimation may be unable to address this problem if it is sufficiently difficult to determine which states are actually engaged in crisis bargaining.

We illustrate our argument in the context of one important and prominent example: military arming. Slantchev (2005, 2011) argues that states can often deter aggression through military preparations for war. Yet the empirical research that finds any consistent relationship typically conclude that arming leads to war.[7] Extant research is insufficient to allow us to determine whether this inconsistency invalidates the theoretical expectation because, as we demonstrate below, even if our theoretical models were *exactly* right about how arming affects crisis bargaining, the typical research design would very likely fail to identify this relationship. For the sake of argument, we assume that the theoretical model we analyze perfectly characterizes crisis bargaining. Then, using simulated data generated directly from

[6]By proxy variables we mean observable variables that correlate with the true omitted variable.

[7]See Colaresi and Thompson (2005), Gibler, Rider, and Hutchison (2005), Sample (1997), Senese and Vasquez (2005), and Senese and Vasqeuz (2008). However, see also Diehl and Crescenzi (1998).

this model, we assess the prospects for identifying the true relationship between arming and conflict.

We begin by presenting a straightforward extension of the canonical ultimatum crisis bargaining model, in which a Target can publicly invest in armaments prior to a Challenger issuing a demand. The equilibria of this model indicate that the Target can generally prevent war by arming. We then construct a simulated data set designed to mirror those typically employed in quantitative research. In doing so, we make three key assumptions: 1) that our bargaining model perfectly describes the data-generating process for those states engaged in crisis bargaining, but 2) that most pairs of states are *not* engaged in crisis bargaining, and 3) that as researchers we cannot observe which states are engaged in crisis bargaining, although observable variables might serve as proxies. We demonstrate, unsurprisingly, that a naïve research design will falsely identify a positive relationship between arming and conflict.[8] We then evaluate three alternate approaches: matching, instrumental variables, and statistical backwards induction.

Our results indicate that the use of instrumental variables can uncover the true negative relationship between arming and conflict, even when using a less appropriate dependent variable (i.e., levels of conflict short of war). In contrast, after matching on observables, we often estimate a positive relationship between arming and conflict. This outcome is particularly likely when the dependent variable reflects low-level conflicts, as is virtually always the case in applied research due to the rarity of full-scale war. Finally, a simplistic application of statistical backwards induction (SBI), one that follows closely the example provided by Bas, Signorino, and Walker (2008), fails to identify the true negative relationship.[9] While this result argues against treating the simple example presented in Bas, Signorino, and Walker

[8]Note that the only possible outcomes in our theoretical model are peace and war. However, much of the quantitative literature tests hypotheses about the causes of war using lower levels of conflict without justifying the implicit assumption that the causes of the former and the latter are the same. We therefore include measures of both in our simulated data, and use the word "conflict" to refer to both.

[9]However, we demonstrate that this failure is not strictly due to the mismatch between the extensive form underlying our application of SBI and that of the theoretical model we used to generate the data, since we would be able to correctly identify a negative relationship were we able to estimate the model solely for observations engaged in crisis bargaining.

(2008) as a one-size-fits-all approach, we stress that this does not detract from the value of truly *theoretically informed* structural estimators, which we do not analyze here.

It is difficult to overstate the challenges we face if we cannot observe those factors that best account for systematic differences in the baseline risk of exposure to the outcome of interest, particularly if only crude proxies for such factors are available. Although some of the tools we analyze prove capable of correctly identifying a negative relationship between arming and conflict under certain circumstances, several features of our Monte Carlo simulations *underestimate* the challenges scholars are likely to encounter in practice. For example, we have assumed that the true relationship is strictly monotonic, that all observable variables are measured without any error, that no significant multi-collinearity exists between the variables, and that our theoretical model not only isolates an important element of the data-generating process but in fact fully characterizes the path to war.[10] We sincerely hope that few scholars would be prepared to make such optimistic assumptions when they work with observational data.

Our results suggest a number of unhappy conclusions. The common practice of estimating binary logit models with measures of conflict short of war as the dependent variable appears nearly certain to mischaracterize the causes of war. Matching methods are likely not the best answer to the problems we have identified. It may be worthwhile to search out valid instruments, though we are cognizant of well known limitations to their use.[11] Finally, we strongly caution against uncritically employing the simplest possible approach to structural estimation without considering whether the underlying theoretical model is appropriate.

Taken together, our results indicate that purely inductive inquiry, or attempts to "let the data speak," will often produce incorrect inferences. In particular, scholars should be very wary about drawing inferences about the impact of behaviors chosen by actors who we have theoretical reason to believe face different risks of exposure to the outcome of interest than do other actors. Whether such concerns apply to a given topic of inquiry is, we believe, best

[10] In other words, we have assumed our model is not in fact a model at all.

[11] See, for example, Bound, Jaeger, and Baker (1995) on the problem of weak instruments.

determined with the assistance of well-developed theory.

2. THE MODEL

We now introduce the model that underlies our simulated data set. The game begins with Nature selecting the Target's military capability, m_2, and revealing it only to the Target. The Target is relatively weak with probability w, in which case Nature sets $m_2 = \underline{m}_2$. The Target is relatively strong with probability $1-w$, in which case Nature sets $m_2 = \overline{m}_2$, where $0 < \underline{m}_2 < \overline{m}_2$. The Target then chooses whether to arm or not. If the Target arms, it incurs cost $\kappa > 0$ and its military capability, m_2, increases by $\alpha > 0$.

After observing the Target's decision, the Challenger demands $x \in [0, 1]$, which the Target may either accept or reject. If the Target accepts, the good is divided accordingly, and the Challenger receives x while the Target receives either $1 - x$ or $1 - x - \kappa$, depending upon whether the Target chose to arm.

If the Target rejects the Challenger's demand, a war occurs. The Challenger wins the war with probability p. We assume that the side that wins will choose to keep the entire value of the good in dispute, allocating nothing to the loser. Moreover, we assume each side incurs some loss of utility associated with incurring the costs of war, denoted $c_1 \in (0, 1]$ and $c_2 \in (0, 1]$ for the Challenger and Target, respectively. Therefore, in the event of war, the Challenger receives $p(1)+(1-p)(0)-c_1 = p-c_1$ and the Target receives $p(0)+(1-p)(1)-c_2 = 1-p-c_2$ (or $1-p-c_2-\kappa$, if the Target chose to arm).

The precise value of p, the probability with which the Challenger defeats the Target, is a function of m_1, m_2, and, if the Target armed, α, where $m_1 > 0$ is the Challenger's military capability. Specifically, if the Target is weak and chose not to arm, $p = \dfrac{m_1}{m_1 + \underline{m}_2} \equiv \overline{p}$. If the Target is weak and the Target armed, $p = \dfrac{m_1}{m_1 + \underline{m}_2 + \alpha} \equiv \hat{p}$. If the Target is strong and chose not to arm, $p = \dfrac{m_1}{m_1 + \overline{m}_2} \equiv \tilde{p}$. Finally, if the Target is strong and armed, $p = \dfrac{m_1}{m_1 + \overline{m}_2 + \alpha} \equiv \underline{p}$. By assumption, $\alpha > 0$ and $\underline{m}_2 < \overline{m}_2$, therefore $\underline{p}$ must be the

smallest value of p and $\overline{p}$ must be the largest.[12]

2.1. *Equilibria*

Since we assume the Challenger is uninformed about the Target's type, we solve the model for perfect Bayesian equilibria, where players' strategies must be sequentially rational and their beliefs weakly consistent with Bayes' Rule. Thus, the Challenger's beliefs about the Target's type must incorporate the information revealed, if any, by the Target's decision to arm or not. The Challenger's strategies must be optimal, both on and off the equilibrium path, given the Challenger's posterior beliefs.[13] There are eight such equilibria in pure strategies, all of which are pooling equilibria. There are four equilibria in which both the strong and weak type of Target pool on not arming and four in which both the strong and weak type of Target pool on arming.[14]

Both when the Target arms and when they do not, there exist equilibria in which the Challenger selects a relatively large value of x, denoted $\overline{x}$, which the Target accepts if and only if the Target's military capability is relatively small, as well as equilibria where the Challenger selects a relatively small value of x, denoted $\underline{x}$, which the Target accepts regardless of type. Because all the equilibria are pooling equilibria, the Target's decision to arm does not influence the Challenger's beliefs. Yet arming still plays a critical role, because it influences the conditions under which the Challenger risks war. The conditions under which the Challenger risks war are more restrictive when the Target arms than when the Target does not arm.

Formally, when the Target arms, the Challenger sets $x = \overline{x}_A$ when $w' > \dfrac{c_1 + c_2}{\hat{p} - \underline{p} + c_1 + c_2} \equiv \overline{w}$, and $x = \underline{x}_A$ when $w' \leq \overline{w}$, where w' denotes the Challenger's posterior belief that the Target is weak given that the Target armed. If the Target does not arm, the Challenger sets

[12]The ordering of $\tilde{p}$ and $\hat{p}$ depends upon whether $\alpha > \overline{m}_2 - \underline{m}_2$.

[13]The Target knows their own military capabilities and can infer the Challenger's beliefs, so the Target's strategies are conditioned on the Challenger's optimal strategies.

[14]Proofs can be found in Appendix A.

$x = \overline{x}_N$ when $w'' > \dfrac{c_1 + c_2}{\overline{p} - \tilde{p} + c_1 + c_2} \equiv \underline{w}$, and sets $x = \underline{x}_N$ when $w'' \leq \underline{w}$, where w'' denotes the Challenger's posterior belief that the Target chose not to arm. Note that either w' or w'' will simply equal w. That is, one of the Challenger's posterior beliefs must always mirror the Challenger's prior belief, given that all our equilibria are pooling equilibria.

Also note that the relatively small ($\underline{x}_A$) and relatively large ($\overline{x}_A$) demands that the Challenger makes after observing the Target arm are distinct from the relatively small ($\underline{x}_N$) and relatively large ($\overline{x}_N$) demands that the Challenger makes if the Target does not arm. This accounts for the difference between $\underline{w}$ and $\overline{w}$. But despite this difference, it remains true that the Target accepts relatively large demands if and only if he is relatively weak, and accepts relatively small demands regardless of his type. The key result here is that $\overline{w} > \underline{w}$, and thus when $\underline{w} \leq w < \overline{w}$, the Challenger makes a demand that carries a risk of war if and only if the Target chooses not to arm.[15]

Table 1 summarizes the equilibria as a function of the Target's decision to arm and the Challenger's prior belief that the Target is weak. Note that there are conditions under which the probability of war is unaffected by the Target's choice of whether to arm or not. Specifically, when $w' \leq \overline{w}$ and $w'' \leq \underline{w}$ or when $w' > \overline{w}$ and $w'' > \underline{w}$. In the former case, war is not expected to occur either way, while in the latter case, the Challenger will risk war regardless of whether the Target arms or not. In contrast, when $w' \leq \overline{w}$ and $w'' > \underline{w}$, the Challenger risks war if and only if the Target chooses not to arm. In such cases, we would conclude that arming strictly prevents war. That being said, when $w' > \overline{w}$ and $w'' \leq \underline{w}$, the Challenger risks war if and only if the Target does not arm. In such cases, arming appears to *promote* war.

[Table 1 about here]

However, unless we make some unusual assumptions about the beliefs held by the Chal-

[15] *Proof.* If $\overline{p} - \tilde{p} > \hat{p} - \underline{p}$, it must be true that $\overline{w} > \underline{w}$, since $\underline{w}$ has a larger denominator. Substituting in the values of p, we obtain $\frac{m_1}{m_1+\underline{m}_2} - \frac{m_1}{m_1+\overline{m}_2} > \frac{m_1}{m_1+\underline{m}_2+\alpha} - \frac{m_1}{m_1+\overline{m}_2+\alpha}$. Since the right hand side of the inequality is identical to the left hand side except that the denominators of both terms include α, pushing both fractions on the right hand side arbitrarily closer to 0, the difference between the two terms on the left must be larger than the difference between the two terms on the right. □

lenger with respect to events that never occur in equilibrium, we can nonetheless conclude that arming *more often than not* prevents war within our model. Suppose, for example, that the Challenger's off the equilibrium path belief matches the Challenger's prior belief. Since $\underline{w} < \overline{w}$, it is impossible for $w' > \overline{w}$ and $w'' \leq \underline{w}$ to simultaneously hold. If we relax this assumption and allow off the equilibrium path beliefs to take on any value, we need only assume that they are distributed similarly to the Challenger's prior beliefs in order to conclude that the conditions under which arming strictly reduces the probability of war will more often be met than the conditions under which arming promotes war. As we discuss below, this is precisely what we assume in our simulated data set.

If, and *only* if, both types of Target arm, and the Challenger's prior belief that the Target is weak is sufficiently large that the Challenger is willing to risk war by issuing a relatively large demand despite this, but the Challenger would believe that the Target is far less likely to be weak if the Target had chosen not to arm, then we would conclude that arming in fact promotes war. It is unclear to us why the Challenger would be expected to hold such a belief. There is no sense in which the stronger Target would face a greater incentive to deviate from the equilibrium strategy of arming.

Nonetheless, in our simulated data set, we allow the Challenger's out of equilibrium belief to take on any value with equal probability, and we still conclude that arming is typically associated with a reduced probability of war. Therefore, we do allow for the possibility that, for *some* observations in which the equilibrium probability of war is non-zero, had the Target deviated from the equilibrium strategy and chosen not to arm, the probability of war would be zero. For this reason, we take care to stress throughout that the true relationship between arming and war in our simulated data is that, *on average*, arming reduces the probability of war, rather than claiming that this is always the case.

Readers who are not persuaded by our treatment of out of equilibrium beliefs should note that, irrespective of whether such assumptions would be warranted if our goal was to use the model to generate observable implications that we then sought to test with observational

data, given the way we set up our Monte Carlo simulations, we must nonetheless conclude that we are being misled by statistical results that fail to identify that, *within our simulated data set*, arming is generally associated with a reduced probability of war.

3. IMPLICATIONS

The equilibria discussed above exhibit several important properties that are useful for our Monte Carlo simulations, where we treat the model as the complete data-generating process (DGP). Note, however, that we do so not because we believe this to be plausible, but rather because it illustrates the problem more clearly.[16]

First, the model generates equilibria in which war occurs despite the Target choosing to arm. Since we evaluate a probabilistic relationship between arming and conflict, the absence of such equilibria could introduce the zero-likelihood problem (Morton 1999). Second, the use of matching is more straightforward when the treatment is binary. Treating the decision to arm as binary in the formal model allows us to use a binary treatment without compromising the connection between the model and the simulated data. Finally, while there are eight different equilibria, each one occupies a unique subset of the parameter space. Thus, we do not need to make arbitrary decisions about equilibrium selection.

It is useful to express the implications of the model more precisely using the potential outcomes framework.[17] Let Y_{i1} be a binary indicator for whether observation i experiences conflict given that the Target armed (is treated) and Y_{i0} be a binary indicator for whether i experiences conflict given that the Target did not arm (is untreated). The fundamental problem of causal inference is that for each i, only one of these outcomes is observed. If a state arms, we cannot know with any certainty what would have happened to that state had it not armed. To do so, we would need to be able to observe a counterfactual world.

[16] If the model were assumed to only partially describe the DGP, it would be less surprising for traditional methods to fail to uncover the relationship described by the model.

[17] See Sekhon (2008) for an overview of this framework.

All we can do is compare the outcomes of those states that did not arm to the outcomes of those that did and hope that any differences we observe in the outcomes, after attempting to correct for other differences between the states, are due to the fact that some armed and some did not.

The average treatment effect (ATE) of arming on conflict is: $\tau = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0)$, where $T_i = 1$ indicates the treatment regime (the Target armed) and $T_i = 0$ indicates the control regime (the Target did not arm). Unfortunately, provided the observations for T_i differ in ways relevant to the probability of conflict, we might identify a positive ATE even if the average observation in the treatment regime would have been more likely to experience conflict were it instead in the control regime, and the average observation in the control regime less likely to experience conflict were it in the treatment regime.

We focus on identifying the average treatment effect for the treated, or ATT, independently of the average treatment effect for the control, or ATC, where ATT is denoted τ_t: $\tau|(T = 1) = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1)$ and ATC is denoted τ_c: $\tau|(T = 0) = E(Y_{i1}|T_i = 0) - E(Y_{i0}|T_i = 0)$. That is, we focus throughout on whether observations in which the Target did in fact arm would have experienced a lower probability of conflict had the Target instead chosen not to arm. This is related to, but distinct from, the question of whether the probability of war found among those observations in which the Target did not arm would have been higher had the Target instead armed. We focus on the former more so than the latter because although the effects are substantively similar in both cases, the ATT should be larger, and thus, the failure to properly identify it would be more striking than the failure to properly identify the ATC would be.[18]

Consider the following example. Suppose 120 civil wars occurred over some time period and that the peace that followed appears to be relatively stable in half of these cases, and relatively fragile in the other half. Critically, let us suppose that the differences between these

[18]There will be more cases where the probability of war is precisely zero in the control group than in the treatment group because treatment status is itself correlated with whether the states are engaged in crisis bargaining. This will attenuate the relationship between arming and war in the control group.

two cases are observable, and that actors are aware of the systematic difference between the two. Now, imagine that, left alone, only one-third of the stable cases will see a return to conflict, whereas the more fragile cases will experience a second war two-thirds of the time.

Now suppose, for the sake of argument, that peacekeeping always has a stabilizing effect, and that this effect is even greater when the situation is fragile. That is, imagine that the probability of conflict recurrence among the stable cases would, on average, be cut by 25% should a peacekeeping operation be deployed, while the fragile cases would see the risk of return to warfare cut by 50%. Finally, imagine that peacekeeping operations are deployed to all of the cases where peace appears fragile, but none of those where it is more stable.

Given the assumptions we have made, we would expect to observe 20 new wars among the 60 states that did not receive peacekeeping operations ($\frac{1}{3} \cdot 60 = 20$) and 20 new wars among the 60 states that received peacekeeping operations ($\frac{1}{2} \cdot \frac{2}{3} \cdot 60 = 20$). If we were to compare the rate of recurrence among those states with peacekeeping operations to those without, we would observe no difference and conclude that peacekeeping was ineffective *even though we have already assumed that it is.*

The assumptions we have made indicate that $E(Y_{i1}|T = 1)$, or the number of civil wars we actually expect to see among those states that received peacekeeping operations, is 20, while $E(Y_{i0}|T = 1)$, or the expected number of civil wars that the states who in fact received peacekeeping operations would have experienced had they never received peacekeeping operations, is 40 ($\frac{2}{3} \cdot 60 = 40$). Since $20 - 40$ is -20, the average treatment effect among the treated, τ_t, is negative. Further, since $E(Y_{i1}|T = 0)$, or the number of civil wars that we would have expected to occur among those states that did not in fact receive peacekeeping operations had they in fact received peacekeeping operations, is 15 ($\frac{3}{4} \cdot \frac{1}{3} \cdot 60 = 15$), and $E(Y_{i0}|T = 0)$, or the number of civil wars that we are actually expected to occur among those states that did not receive peacekeeping operations, is 20, and $15 - 20$ is -5, and the average treatment effect among the control regime, or τ_c, is also negative. In fact, there were 115 civil wars from 1945 to 2004, and 39% of those that received a peacekeeping operation

saw a return to conflict versus 42% of those that did not. Yet once we account for observable differences between states that received peacekeeping operations and those that do not, peacekeeping indeed appears to prevent future civil wars (Fortna 2004, Gilligan and Sergenti 2008).

The critical point here, though, is that the differences between cases where peace was stable and those where it was fragile are observable, not only to actors within the international system, but to scholars as well. Yet, as we argued above, there may be reasons to suspect that it is difficult to determine which states are engaged in crisis bargaining. Since crisis bargaining is both a predictor of war and a predictor of arming, we have reason to doubt that extant empirical analysis of the relationship between arming and war is less likely to have identified the true relationship. Note that we are not claiming to know that arming prevents war. It is entirely possible that the theoretical models that suggest that it does are hopelessly flawed. Our argument is simply that *even if it were true* that arming prevents war, we might falsely conclude that it does not – and that this is a much harder problem to solve than is commonly appreciated.

As discussed above, in Table 1 we see that when $w' \leq \overline{w}$ and $w'' \leq \underline{w}$ or when $w' > \overline{w}$ and $w'' > \underline{w}$, the probability of war is independent of the Target's decision to arm. In the language of potential outcomes, that means that there are cases for which $E(Y_{i1}) - E(Y_{i0}) = 0$ and thus $\tau_t = 0$. Naturally, the same will be true for all observations where the states did not engage in crisis bargaining, since the probability of war in such cases is 0, regardless of treatment status.

There also exist cases where $E(Y_{i1}) - E(Y_{i0}) > 0$ as well as cases where $E(Y_{i1}) - E(Y_{i0}) < 0$.[19] That is, as discussed above, it is possible that arming can either prevent or promote war. However, the conditions under which there is a positive probability of war if and only if the Target arms are restrictive. Given the distribution of beliefs for the Challenger that we use in our Monte Carlo simulations (see Appendix B), we can be confident that the *average*

[19]Note that these cases will exist in both the treatment and control groups.

treatment effects amongst both the treatment and control regimes will be negative, and thus, if we fail to identify a negative τ_t and τ_c within our simulated data we will draw an incorrect inference regarding the effect of arming on the likelihood of war.

3.1. *Simulated Data*

We begin by generating the military capabilities for State 1 and State 2 for 100,000 dyads.[20] Specifically, $CAP_1 \sim N(0.5, 0.05)$, $CAP_2^L \sim N(0.5, 0.05)$, and $CAP_2^H = CAP_2^L + \epsilon$, where $\epsilon \sim N(1, 0.1)$. With respect to the bargaining model, CAP_1 represents m_1, CAP_2^L represents $\underline{m}_2$, and CAP_2^H represents $\overline{m}_2$. Next, we generate $ALPHA \sim N(0.5, 0.05)$ to represent α, $KAPPA \sim N(0.05, 0.005)$ for κ, and $COST_i \sim N(0.1, 0.01)$ for $i \in \{1, 2\}$ to represent c_i.

CRISIS is a binary variable indicating whether the two states in the dyad engaged in crisis bargaining, which is true for approximately 10% of the observations (randomly selected). We assume crisis bargaining is perfectly described by the model above.[21] $WAR = 1$ in approximately 19% of the observations where $CRISIS = 1$ in our simulated data. For all other dyads, that is, those where $CRISIS = 0$, war is never observed.

Note that if State 2 arms in order to deter one potential aggressor, their arming decision will be reflected in all dyads containing State 2, since we cannot readily determine which state prompted an increase in State 2's military capabilities. Thus, we create a binary variable ARM, which equals 1 whenever $CRISIS = 1$ and the formal model indicates that State 2 would arm, and also equals 1 when $CRISIS = 0$ with probability 0.25 if $KAPPA$ is less than or equal to its mean value, and with probability 0.1 if $KAPPA$ is above its mean value. In the end, $ARM = 1$ in approximately 41% of the observations when $CRISIS = 1$, and approximately 18% when $CRISIS = 0$.

[20] In those dyads where the states play the crisis bargaining game, State 1 is the Challenger and State 2 is the Target. However, since most dyads in our simulated data do not engage in crisis bargaining, we opt for more general language.

[21] In order to determine equilibrium behavior, we also generated variables representing the Challenger's beliefs. See Appendix B for more details.

IV_1 and IV_2 serve as instruments for ARM. $IV_1 = 1$ with probability 0.75 if $KAPPA \leq 0.05$ and equals 0 otherwise. IV_2 is constructed identically, except the probability is 0.9 instead of 0.75. $IV_1 = 1$ approximately 34% of the time when $ARM = 0$, and approximately 53% of the time when $ARM = 1$. For IV_2, these figures are 40% and 64%, respectively.

Since interstate wars are relatively rare, scholars often test hypotheses about war using measures of lower levels of hostility, such as Militarized Interstate Disputes (MIDs).[22] We create a variable *MID* that equals 1 when $WAR = 1$ but also equals 1 with probability 0.75 if $CRISIS = 1$ and $WAR = 0$, and also with probability 0.05 when $CRISIS = 0$. Approximately 64% of our simulated MIDs reflect actual crises, in the sense that the states played the crisis bargaining game described above.

A few characteristics of our *MID* variable are worth emphasizing. First, a greater percentage of our simulated *MID*s represent wars than is typically true when using the actual MID data. $WAR = 1$ in approximately 16% of our observations when $MID = 1$. In contrast, less than 4% of the MIDs in the data set used by Bennett and Stam (2004) are wars, for example. Our results will therefore *underestimate* the problems associated with testing hypotheses about war using a dependent variable that measures lower levels of hostility, as our simulated *MID* variable is less distinct from *WAR* than is the case of actual MIDs.

Second, as noted above, $CRISIS = 0$ for a substantial number of observations where our simulated *MID* variable equals 1 ($\approx 36\%$). The MID data were designed to minimize the possibility of such cases. According to Jones, Bremer, and Singer (1996, 168), "States do not engage in militarized actions unless they perceive that the issues at stake are important, and ... we believe that militarization is a valid indicator that a dispute is serious." Thus, our simulated *MID* variable may correlate with *CRISIS* to a lesser extent than is the case

[22] See Ghosn, Palmer, and Bremer (2004) for a description of the MID data set and coding rules. Some scholars also focus on "crises", as defined by the International Crisis Behavior (ICB) data set. Note that an ICB crisis entails observable hostile behavior, though typically well below the level of war, and so should not be assumed to be synonymous with engaging in crisis bargaining in the sense used here. As with the MID data, all wars are ICB crises, but not all ICB crises are wars. Although the correlation between MID and ICB crisis is only moderate, they have sufficiently similar properties that the argument developed here should apply to both, though perhaps not to quite the same degree. See Hewitt (2003) for more details.

with respect to the actual data. It is important to note that this too ensures that we are *underestimating* the problems associated with testing hypotheses about the causes of war using the MID data. As *MID* and *CRISIS* become less distinct, we become more likely to estimate a positive relationship between *ARM* and *MID*, given that *CRISIS* predicts *ARM*.

We also create a variable, *RECIP* to indicate whether State 2 reciprocated State 1's hostility, as per the same variable in the MID data. Specifically, we set $RECIP = 1$ with certainty if $WAR = 1$, with probability 0.5 if $CRISIS = 1$, $MID = 1$, and $WAR = 0$, and with probability 0.075 if $CRISIS = 0$, $MID = 1$, and $WAR = 0$. *RECIP* is coded as missing if $MID = 0$. Approximately 43% of our simulated MIDs involve reciprocated hostility.

While we assume that *CRISIS* is unobservable, we generate two dummy variables, $PROX_1$ and $PROX_2$, which we assume are observable and can be used to indicate which dyads are likely to experience conflict. These variables might represent geographic contiguity or rivalry, which have been shown to influence conflict onset.[23] Importantly, neither contiguity nor rivalry need to be considered a *cause* of conflict so much as a strong indicator of which states are likely to engage in crisis bargaining.

We initially set $PROX_1 = 1$ with probability 0.5 if $CRISIS = 1$, and with probability 0.1 if $CRISIS = 0$. $PROX_2 = 1$ with probability 0.3 if $CRISIS = 1$, and with probability 0.1 if $CRISIS = 0$. Thus, $PROX_1 = 1$ approximately 10% of the time when $CRISIS = 0$, and approximately 50% of the time when $CRISIS = 1$ while the corresponding figures for $PROX_2$ are approximately 10% and 30%. We refer to the data set containing these proxy variables as Experiment 1.

Next, we generate a new simulated data set identical to the first one, except that $PROX_1 = 1$ with probability 0.4 when $CRISIS = 1$, and with probability 0.15 when $CRISIS = 0$, while $PROX_2 = 1$ with probability 0.2 when $CRISIS = 1$, and with probability 0.1 when $CRISIS = 0$. Thus, $PROX_1 = 1$ approximately 15% of the time when $CRISIS = 0$, and approximately 40% of the time when $CRISIS = 1$ while $PROX_2 = 1$

[23] See, for example, Bennett and Stam (2004).

approximately 10% of the time when $CRISIS = 0$, and approximately 20% of the time when $CRISIS = 1$. We refer to the data set containing these proxy variables as Experiment 2.

3.2. *Baseline Analysis*

We illustrate the problem by comparing the results from a naïve research design to those obtained after conditioning on the unobservable variable *CRISIS*, which is not possible in practice.[24] For our naïve research design, we estimate binary logits using *MID* and *WAR* as dependent variables. The primary independent variable is *ARM*, although we include the proxy variables from Experiment 1 and Experiment 2 as control variables. We initially include all dyads, as is common in applied research. Next we match on *CRISIS*. For these models, we do not include our proxy variables, since they are no longer necessary in order to isolate the systematic difference between the treatment and control groups.

Table 2 summarizes the results from 100 independently created data sets, where each data set was constructed as described above. We report the mean and standard deviation of the coefficient estimate for *ARM* ($\hat{\beta}$) across the 100 estimated models.[25] We also report the standard error calculated as $\frac{\text{S}}{\sqrt{100}}$ and the rejection rate of $\hat{\beta}$ based on H_0: $\beta = 0$ versus H_a: $\beta \neq 0$ using a two-tailed test and significance level of $p < 0.05$. Note that we use a two-tailed test despite having a theoretical expectation about the direction of the effect because we wish to mimic the analysis an applied researcher would conduct, and there is no consistent empirical evidence indicating that arming prevents war.

[Tables 2 and 3 about here]

[24] Note that *CRISIS* is not the only unobservable variable that systematically differs with treatment status. For example, the cost of arming is typically lower for the treated (those that armed). However, *CRISIS* is a clear confounder, as it is directly related to the likelihood of war, while the cost of arming is not. This is tantamount to matching. However, it tells us nothing about the use of matching in practice, since *CRISIS* is unobservable. Below, when we evaluate matching methods, we focus on observable variables.

[25] For reasons of space, we only report the estimated effect of *ARM*. Unsurprisingly, $PROX_1$ and $PROX_2$ are positive and significant in each of the models in which they are included, that is, when $CRISIS = 0$. More often than not, these variables are not significant, and we would not expect them to be, when they are included in the models where $CRISIS = 1$. Results are available in the web appendix.

The top panel of Table 2 contains the results of our baseline analysis. When the analysis includes all dyads, we *always* estimate a positive relationship between arming and conflict, irrespective of whether conflict is measured using a low-level indicator (*MID*) or *WAR*. This is true for both Experiment 1 and Experiment 2. We also always reject the null hypothesis when analyzing all dyads. When we condition on *CRISIS*, we *always* observe the negative relationship anticipated by the formal model regardless of the dependent variable and for both Experiment 1 and Experiment 2.[26] Additionally, we can *always* reject the null hypothesis when we condition on *CRISIS* and employ *WAR* as the dependent variable, and nearly always when we use *MID* as the dependent variable.

Note that while the choice of dependent variable (*MID* versus *WAR*) does not change the interpretation of the overall relationship between *ARM* and conflict, it does affect the size of the effect. Table 3 shows the predicted probabilities and 95% confidence intervals when $PROX_1 = PROX_2 = 1$, which corresponds to the case where states are most likely to arm, for Experiment 1 and Experiment 2, from which we can estimate $\hat{\tau}_t$.[27] The top panel of Table 3 contains the results of our baseline analysis. When we include all dyads, the change in the size of the effect in each case is larger when the dependent variable is *MID* compared to when it is *WAR*. However, the change in the size of the effect when we condition on *CRISIS* is larger when the dependent variable is *WAR* compared to *MID*. *ARM* decreases the probability of *WAR* by approximately 0.11 (from 0.24 to 0.13) amongst those dyads engaged in crisis bargaining. However, amongst the same dyads, *ARM* only decreases the probability of *MID* by approximately 0.03 (from 0.81 to 0.78).

This baseline analysis illustrates the challenge to correct inference about the relationship between arming and conflict. Despite including relevant control variables, $PROX_1$ and $PROX_2$, we *always* estimate a positive average treatment effect amongst the treated ($\hat{\tau}_t$)

[26]The estimated effect for *ARM* is the same for both Experiment 1 and Experiment 2 because the proxy variables are not included when we only analyze those dyads that engaged in crisis bargaining.

[27]Since we focus on $\hat{\tau}_t$, we set $PROX_1 = PROX_2 = 1$. The results for $PROX_1 = PROX_2 = 0$, which corresponds to the case where states are least likely to arm, from which we can estimate $\hat{\tau}_c$, are available in the web appendix.

when analyzing all dyads. Yet, below, we will see that at least in Experiment 1, $PROX_1$ and $PROX_2$ provide sufficient information to correctly identify a negative $\hat{\tau}_t$. In contrast, once we condition on *CRISIS*, which we cannot do in practice, we are able to correctly identify a negative $\hat{\tau}_t$ using a simple binary logit.

3.3. *Matching*

We now examine whether it is possible to estimate a negative $\hat{\tau}_t$ after matching on $PROX_1$ and $PROX_2$.[28] Above, we implicitly employed exact matching (using *CRISIS*), and correctly identified a negative $\hat{\tau}_t$.[29] However, matching is not guaranteed to reliably uncover treatment effects when important confounders are unobservable since we cannot invoke the assumption of conditional independence (Sekhon 2008).[30] The simulated data violate the conditional independence assumption. However, in practice, researchers can never know for certain whether they have violated this assumption. They can only demonstrate that their matched dataset is balanced with respect to their *observable* covariates and hope for the best.

We restrict our focus to observable variables while employing exact matching, which is the preferred approach to matching when feasible.[31] Exact matching isolates much, but not all, of the systematic difference between the treatment and control groups. In our simulated data, *CRISIS* is 2.8 times as likely to equal 1 in observations where $ARM = 1$ prior to matching, versus 1.46 after matching using the first proxy variable settings and 1.94 using

[28] We focus our discussion on $\hat{\tau}_t$ and not $\hat{\tau}_c$. Analogous models were estimated, when $PROX_1 = PROX_2 = 0$, which corresponds to those cases where *ARM* is least likely to equal 1. We consistently estimated $\hat{\tau}_c$ to be positive and significant. We also estimated models after matching on all possible combinations of $PROX_1$ and $PROX_2$ using the Matching package in R (Sekhon 2011). The primary difference between those results and the ones we present below is that the $\hat{\tau}_t$ is always positive. The $\hat{\tau}_t$ is also smaller than the results we present below when *MID* is the dependent variable. Results available in the web appendix.

[29] Given that *WAR* never occurs when *CRISIS* takes on a value of 0, had we attempted to estimate $\hat{\tau}_c$, we would find that the outcome variable does not vary.

[30] Formally, $\{Y_0, Y_1 \perp\!\!\!\perp T|X\}$, where $\perp\!\!\!\perp$ denotes independence and X is a set of conditioning variables.

[31] Matching based on propensity scores, Mahalanobis Distance, and genetic matching are more often employed in practice, as the curse of dimensionality often renders exact matching impossible, particularly with continuous covariates. See Sekhon (2009) for a discussion of this issue. However, since we only have two observable conditioning variables and both are binary, it is straightforward to implement exact matching by restricting our focus to those dyads where both proxy variables equal 1.

the second proxy variable settings.[32]

The bottom panel of Table 2 summarizes the results after matching on $PROX_1 = PROX_2 = 1$.[33] In the baseline analysis, the primary difference between the models where the dependent variable was *MID* and those where it was *WAR* was in the magnitude of the effect of *ARM*. Here, in both Experiment 1 and Experiment 2, we estimate a positive relationship between *ARM* and conflict when using *MID* as the dependent variable and *always* reject the null hypothesis. When we employ *WAR* as the dependent variable we identify a positive $\hat{\tau}_t$ in Experiment 2 but a negative $\hat{\tau}_t$ in Experiment 1; however, we do not always reject the null hypothesis in either Experiment 1 or Experiment 2. Thus, if we select a more appropriate dependent variable, we might at least avoid drawing the opposite conclusion, although we will still fail to conclude that arming often prevents war.

The middle panel of Table 3 shows the predicted probabilities and 95% confidence intervals when $ARM = 0$ and $ARM = 1$. *ARM* increases the probability of *MID* by approximately 0.17 and 0.19 in Experiment 1 and Experiment 2, respectively, and increases the probability of *WAR* by approximately 0.01 in Experiment 2, but *decreases* the probability of *WAR* by approximately 0.02 in Experiment 1. In the baseline analysis *ARM* decreased the probability of *WAR* by approximately 0.11, nearly six times larger than the effect observed here. Thus, we underestimate the magnitude of the relationship between *ARM* and *WAR* even when the direction of the effect is correct.

These results highlight a common misconception about the inclusion of control variables in linear and generalized linear models, which does not justify the use of the "all else equal" language so common in applied research. Our proxy variables proved capable, at least under some conditions, of leading us to the correct inference when we used them to identify which observations were most comparable rather than including them as controls, as we did in the baseline analysis.

[32]That is, prior to matching, $CRISIS = 1$ in 21% of observations when $ARM = 1$ and 7% when $ARM = 0$. After matching, the corresponding percentages are 80% and 55%, respectively, in Experiment 1 and 58% and 30%, respectively, in Experiment 2.

[33]Note that the sign on $\hat{\beta}$ tells us the direction of $\hat{\tau}_t$ since the only effect of *ARM* here is captured by $\hat{\beta}$.

Note that we have addressed neither the Stable Unit Treatment Value Assumption (SUTVA), nor whether our conditioning variables were determined prior to treatment status. In this case, neither is a concern. But in other cases, they very well might be.[34] Thus, while matching sometimes enabled us to correctly identify a negative $\hat{\tau}_t$, it did not always, and many of the assumptions we have made will often be untenable in practice.

3.4. *Instrumental Variables*

Unlike matching methods, instrumental variables can enable researchers to uncover the true causal effect even when there are unobserved confounders.[35] By construction, our instruments (IV_1 and IV_2) are correlated with ARM but are not directly related to either of our dependent variables (MID and WAR). Moreover, we can be sure that the *average* effect of ARM is non-zero and monotonic. Therefore, we have satisfied the assumptions required for drawing causal inferences using instrumental variables (Angrist and Pischke 2009).[36]

For Experiment 1 and Experiment 2, we estimate four bivariate probit models.[37] We model ARM as a function of the binary instrument and the two proxy variables, and conflict as a function of ARM and the proxy variables. The log likelihood function for each model

[34]SUTVA posits that $Y_{i0}, Y_{i1} \perp\!\!\!\perp T_j \; \forall \; i \neq j$. There is little reason to believe that the Target's ability to deter the Challenger by arming depends upon how many other states in the system arm. It is not clear that the same could be said of other variables of interest to International Relations scholars, such as regime type, possession of nuclear weapons, etc. With respect to matching on pre-treatment variables, there is little reason to believe we face any issues here, particularly if we think the proxy variables refer to geographic factors. However, this requirement is often more demanding in practice.

[35]Provided the instrument is not systematically related to the confounders (Angrist and Pischke 2009).

[36]More formally, there are four assumptions that must be met: Independence of the instrument, or $\{Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}\} \perp\!\!\!\perp Z_i$, where D_{1i} and D_{0i} are potential treatment statuses and Z_i is a binary indicator for the instrument; Exclusion, or $Y_i(d, 0) = Y_i(d, 1) \equiv Y_{di}$ for $d \in \{0, 1\}$; First Stage, or $E[D_{1i} - D_{0i}] \neq 0$; and Monotonicity, which requires $D_{1i} - D_{0i} \geq 0 \; \forall \; i$ or $D_{1i} - D_{0i} \leq 0 \; \forall \; i$.

[37]Bivariate probit models allow researchers to simultaneously estimate two equations, a feature which is useful when the outcome variable of interest is binary as is the endogenous regressor (Angrist and Pischke 2009, 198–205). In International Relations, the most frequent use of this estimator is not associated with instrumental variables, but selection effects, that is, for models where Y_2 takes on non-missing values only when $Y_1 = 1$ (e.g., Reed (2000)). See also Xiang (2010), whose approach is very similar to ours, although he does not discuss his use of bivariate probit in the context of instrumental variables. We used Zelig to estimate our bivariate probit models (Imai, King, and Lau 2007, 2008).

is:

$$\sum \left[Y_i \ln \Phi_b \left(\frac{X_i\beta_0 + ARM_i\beta_1}{\sigma_\epsilon}, \frac{X_i\gamma_0 + IV_i\gamma_1}{\sigma_\nu}; \rho_{\epsilon\nu} \right) + (1 - Y_i) \ln \left[1 - \Phi_b \left(\frac{X_i\beta_0 + ARM_i\beta_1}{\sigma_\epsilon}, \frac{X_i\gamma_0 + IV_i\gamma_1}{\sigma_\nu}; \rho_{\epsilon\nu} \right) \right] \right], \quad (1)$$

where Y_i is either *MID* or *WAR*, $X_i = PROX_1 + PROX_2$, β_0 and β_1 are coefficients to be estimated in the second stage, γ_0 and γ_1 are coefficients to be estimated in the first stage, ϵ is the error term for the second stage, ν is the error term for the first stage, and $\Phi_b(\cdot, \cdot, \rho_{\epsilon\nu})$ is the bivariate normal cumulative distribution function with correlation coefficient $\rho_{\epsilon\nu}$.

We are primarily interested in the ATT, or $\hat{\tau}_t = E[Y_{1i} - Y_{0i}|IV_i = 1]$. As Angrist and Pischke (2009, 200–201) discuss:

$$\hat{\tau}_t = E \left[\frac{\Phi_b \left(\frac{X_i\hat{\beta}_0 + \hat{\beta}_1}{\sigma_\epsilon}, \frac{X_i\hat{\gamma}_0 + IV_i\hat{\gamma}_1}{\sigma_\nu}; \rho_{\epsilon\nu} \right) - \Phi_b \left(\frac{X_i\hat{\beta}_0}{\sigma_\epsilon}, \frac{X_i\hat{\gamma}_0 + IV_i\hat{\gamma}_1}{\sigma_\nu}; \rho_{\epsilon\nu} \right)}{\Phi \left(\frac{X_i\hat{\gamma}_0 + IV_i\hat{\gamma}_1}{\sigma_\nu} \right)} \right]. \quad (2)$$

Equation 2 defines $\hat{\tau}_t$ as the difference in the expected probability of conflict given that the Target armed and the expected probability of conflict given that the Target did not arm divided by the expected probability that the Target arms. Thus, the ATT can be identified readily from the predicted probabilities, which are presented in Table 4. We calculate $\hat{\tau}_t$ when setting $IV_1 = PROX_1 = PROX_2 = 1$, thereby maximizing the probability of treatment.[38]

[Table 4 about here]

For Experiment 1, when Y_i is given by *MID*, we calculate $\hat{\tau}_t \approx -0.08$.[39] We calculate $\hat{\tau}_t \approx -0.03$ for *WAR*. For Experiment 2, the corresponding figures are approximately -0.10 and

[38] We have also calculated predicted probabilities where $PROX_1 = PROX_2 = 0$. Results available in the web appendix.

[39] Where $\hat{\tau}_t = \frac{Pr(Y_1 = 1, Y_2 = 1|ARM = 1) - Pr(Y_1 = 0, Y_2 = 1|ARM = 0)}{Pr(Y_1 = 1, Y_2 = 1|ARM = 1) + Pr(Y_1 = 1, Y_2 = 0|ARM = 1)} = \frac{0.22 - 0.25}{0.22 + 0.18} \approx -0.08$.

-0.03, respectively. Thus, $\hat{\tau}_t$ is consistently negative, although surprisingly, the estimated effect is larger for *MID* than *WAR*.[40]

These results are only somewhat encouraging. We implicitly made some strong assumptions.[41] We did not discuss local treatment effects nor how we might identify an instrument for arming.[42] Such concerns are more difficult to ignore when analyzing real data.[43]

3.5. *Statistical Backwards Induction*

Finally, we assess the prospects for identifying the negative relationship between arming and conflict using structural estimation, specifically statistical backwards induction (SBI). SBI has the advantage of being simpler to employ than most structural estimators derived explicitly from game-theoretic models, with a minimal loss of efficiency.[44]

We assume a similar game structure to that presented in Figure 1 of Bas, Signorino, and Walker (2008). Specifically, we implement SBI assuming the game begins with a decision by the Challenger to accept the status quo or initiate a crisis, in which case the Target decides to concede or to reciprocate, where reciprocation produces a conflict.

There are two important differences between this extensive form and the one in our crisis bargaining model. First, when using SBI, the good in dispute is treated as indivisible.[45] It

[40]The results when we replace IV_1 with IV_2 are very similar and available in the web appendix.

[41]Not the least of which being that we identified the maximum of the likelihood function. Freedman and Sekhon (2010) note that canned bivariate probit routines in statistical software packages often fail to identify global maxima.

[42]The causal effects estimated using instruments are called local treatment effects because they depend upon the instrument. The estimated effect pertains only to those observations who would enter the treatment regime if they also receive the instrument but would remain in control regime otherwise (compliers). If any members of the population would enter the treatment regime only if they *do not* receive the instrument (non-compliers), our inferences will be invalid. Our theoretical model gives us no reason to expect any of the observations to be non-compliers, nor any reason to believe the effects for compliers will differ systematically from those for the remainder of the population.

[43]See Sovey and Green (2011) for a nice checklist of the assumptions made in instrumental variables estimation and the evidence that should be provided in defense of the assumptions.

[44]See also Signorino (1999, 2003) and Signorino and Yilmaz (2003).

[45]We are not aware of any practical means for measuring the demands states make of one another. As demonstrated in the baseline analysis, this would not necessarily hinder our ability to empirically evaluate comparative static predictions relating the occurrence of conflict to the Target's decision to arm provided we could accurately determine which states are engaged in crisis bargaining.

is difficult to determine how important this difference is. For example, it is possible that in practice the decision to initiate a MID approximates the decision to propose a division of the good in dispute that runs a greater risk of rejection. Unfortunately, few scholars have addressed the question of what MIDs actually represent with respect to bargaining outcomes.

Second, we do not treat the decision to arm as endogenous in the model that will be tested using SBI. Rather, our application of SBI treats this decision as given and seeks to evaluate its influence on the player's subsequent decisions by assuming that each player's payoff for conflict depends upon whether the Target armed previously. This is very likely an important difference. We wish to stress that we do not intend the analysis we conduct to serve as a test of the power of properly applied and theoretically appropriate structural estimators. A growing literature on this topic strongly indicates that such applications are powerful and we hope to see their use become more common. Rather, our goal is to demonstrate the pitfalls associated with conducting empirical analysis that is divorced from prevailing theories, and that this problem may not be solved by more sophisticated methods. In short, we believe that many scholars might assume this game structure because of the ease with which it can be implemented thanks to the sample code provided by Bas, Signorino, and Walker (2008).

Naturally, it would be more appropriate to analyze escalation to war within the subset of observations we can be confident are engaged in crisis bargaining, even if we continued to treat the decision to arm as exogenous. In fact, we demonstrate below that if it were possible to observe *CRISIS*, one could uncover the correct inference even using the game structure that underlies our application of SBI.[46]

We specify the utilities at each terminal node as follows.[47] As Bas, Signorino, and Walker (2008) point out, it is necessary to constrain one of the utilities to be 0 in order to identify the model. Following their suggestion, we assign a value of 0 to the Challenger's utility for the status quo, as well as for the Target's utility for conceding (see Figure 2 in Bas, Signorino,

[46] We also wish to stress that it would be straightforward to adapt the code provided by Bas, Signorino, and Walker (2008) to allow for additional moves (such as the decision to arm). Our analysis cannot speak to the potential of more sophisticated structural estimators.

[47] See the web appendix for the extensive form of our game.

and Walker (2008)). The remaining utilities (the payoffs for each player in the event that the Target reciprocates, and the Challenger's utility for having the Target concede) consist of three parts: a coefficient, a set of regressors, and an error term. More formally, let $X_{cC}\beta_{cC}$ represent the Challenger's utility for having the Target concede, $X_{cR}\beta_{cR}$ the Challenger's utility for the Target reciprocating, and $X_{tR}\beta_{tR}$ the Target's utility for reciprocating. For the sake of simplicity, let $X_{cC} = 1$, $X_{cR0} = 1$, $X_{cR1} = (1 - ARM)$, $X_{cR2} = CAP_1$, $X_{tR0} = 1$, $X_{tR1} = ARM$, $X_{tR2} = PROX_1$, and $X_{tR3} = PROX_2$.[48]

We require the player's choices to be probabilistic. Therefore, we assume that the Target reciprocates if $X_{tR}\beta_{tR} + \epsilon_t > 0$ and the Challenger initiates a MID if $(1 - p_r)(X_{cC}\beta_{cC}) + (p_r)(X_{cR}\beta_{cR} + \epsilon_c) > 0$, where p_r is the probability that the Target reciprocates. Here, ϵ_c and ϵ_t can be interpreted as either agent error or private information. As long as we assume that we as analysts cannot know the true value of ϵ_c and ϵ_t, while the Challenger knows the value of ϵ_c but not ϵ_t and the Target knows the value of ϵ_t but not ϵ_c, we can remain agnostic as to whether these terms reflect agent error, and thus suboptimal or non-Nash play, or private information held by actors who behave optimally given the information they have.[49] We also assume ϵ_c and ϵ_t are independently distributed Type I Extreme Value, such that the choice probabilities are logit probabilities with uncorrelated error terms.[50]

Note that the effect of ARM is ambiguous. By construction, ARM has two important effects. Most obviously, setting $X_{cR1} = (1 - ARM)$ ensures that the Challenger's utility for reciprocation is decreasing in ARM. However, ARM is also found in X_{tR}. Provided $\beta_{tR1} > 0$, this effect increases p_r. Whether an increase in p_r increases or decreases the probability that

[48] We include CAP_1 in X_{cR} but do not include the Target's capabilities (CAP_2) in X_{tR} because we have assumed the Target alone knows the true value of their capabilities. Moreover, our primary interest is in the net influence of ARM on MID rather than the influence of pre-existing capabilities.

[49] The distinction is immaterial here because the Target's behavior does not depend upon their beliefs about the Challenger's willingness to initiate a dispute. If the Target's decision depended upon ϵ_c, assuming the ϵ terms reflected agent error would in fact differ from assuming they reflect private information since the Target's updated belief about ϵ_c after observing the Challenger's initial decision would become important. See Lewis and Schultz (2003), Wand (2006), and Whang (2010) for examples of such estimators. Yet, in the current formulation, the Target need only know their own payoffs, and the Challenger must be uncertain about the Target's strategy.

[50] See Bas, Signorino, and Walker (2008, 27).

$MID = 1$ is unclear, since $X_{cC}\beta_{cC}$ may be either larger or smaller than $X_{cR}\beta_{cR}$.[51]

We estimate $\hat{p}_c$ and $\hat{p}_r$ as follows. We begin by estimating a logit with *RECIP* as the dependent variable. We evaluate the probability of reciprocation using X_{tR}, where X_{tR} is specified as above. The predicted probabilities generated from this model produce $\hat{p}_r$. Next, we construct $Z_{cC} = (1 - \hat{p}_r)X_{cC}$ and $Z_{cR} = \hat{p}_r X_{cR}$, where X_{cC} and X_{cR} are specified as discussed above. We then estimate a second logit with *MID* as the dependent variable and Z_{cC} and Z_{cR} as the independent variables. This ensures that the coefficient estimates for Z_{cC} and Z_{cR} will be $\hat{\beta}_{cC}$ and $\hat{\beta}_{cR}$, providing estimates of the β terms identified above.[52] Since the regressors in the second logit are predicted probabilities, we must calculate bootstrapped standard errors in order to avoid biased estimates for the standard errors of $\hat{\beta}_{cC}$ and $\hat{\beta}_{cR}$.[53]

The bottom panel of Table 3 shows the predicted probabilities of *MID* and 95% confidence intervals when $PROX_1 = PROX_2 = 1$, which corresponds to the case where states are most likely to arm, for Experiment 1 and Experiment 2.[54] We held CAP_1 at its mean value of 0.5.[55] In Experiment 1, *ARM* increases the probability of a MID ($\hat{p}_c$) by approximately 0.20. In Experiment 2, *ARM* increases the probability of *MID* by approximately 0.19. These results indicate that SBI, using a specification similar to that provided by Bas, Signorino, and Walker (2008), is unlikely to correctly identify the relationship between *ARM* and conflict. However, we wish to be clear about what the source of the problem here is, which is not a fundamental flaw in the logic underlying SBI.

To illustrate this, we follow the same procedure as above, with two small changes. First, we focus only on observations where $CRISIS = 1$. Second, because we have little reason to expect $PROX_1$ or $PROX_2$ to capture important variation within this subset of observations, we drop X_{tR2} and X_{tR3} from X_{tR}. Otherwise, we proceed precisely as above. The bottom

[51] The probability that $MID = 1$ decreases in p_r provided $\beta_{cC} > \beta_{cR0} + \beta_{cR1}(1 - ARM) + \beta_{cR2}CAP_1$.

[52] We use $\hat{\beta}_{cC}$ and $\hat{\beta}_{cR}$ here as shorthand for the individual coefficient estimates on each of the X values that comprise X_{cC} and X_{cR}.

[53] We use a nonparametric bootstrap with 500 bootstrap iterations.

[54] We set $PROX_1 = PROX_2 = 1$ to estimate $\hat{\tau}_t$. Results when $PROX_1 = PROX_2 = 0$ are available in the web appendix.

[55] Results when CAP_1 was held at its minimum and maximum values are available in the web appendix.

panel of Table 3 also shows the predicted probabilities of *MID* and 95% confidence intervals when $CRISIS = 1$. These results show that arming decreases the probability of a *MID* by approximately 0.02 in both Experiment 1 and Experiment 2. This reinforces our claim that the problem is that *CRISIS* is unobservable rather than a fundamental shortcoming of SBI.

We also wish to stress that we have focused on the simplest possible application of SBI rather than one that is more theoretically appropriate (i.e., one where we treat *ARM* as endogenous, as in the bargaining model that we presented above). Our goal is to illustrate the consequences of failing to think seriously about the implications of prevailing theoretical arguments, as an uncritical adaptation of the code provided by Bas, Signorino, and Walker (2008) does, rather than to assess the performance of structural estimators that are more theoretically motivated, which we have every reason to believe would perform much better.

4. SUMMARY OF RESULTS

Before concluding, we briefly summarize our results. Throughout, we have focused our discussion on $\hat{\tau}_t$, the average treatment effect amongst the treated. Put differently, we have focused on the difference between the observed likelihood of conflict for those observations where the Target armed and the counterfactual likelihood of conflict that would have been observed for those observations had the Target not armed. Figure 1 presents the estimated τ_t for our baseline analysis and each of the three empirical methods discussed above.

[Figure 1 about here]

Overall, the results from the bivariate probit models with instrumental variables performed best, consistently estimating a negative τ_t regardless of the choice of dependent variable or the quality of the proxy variables. Our simplistic approach to statistical backwards induction fared worst of all. Matching on observables uncovered a modest negative effect for Experiment 1 when the dependent variable was *WAR*.

There are important lessons to be learned from these results. If *CRISIS* was observable, sophisticated methods would not be necessary to correctly infer that, on average, potential Targets are less likely to experience conflict if they arm. The method that performed best was the only one that explicitly modeled the Target's decision to arm. This suggests that scholars who hope to simply "let the data speak" without thinking carefully about how the data were generated in structuring their analysis are likely to draw incorrect inferences.

However, the question of which method will be most appropriate in practice may differ. Instrumental variables may not outperform the other approaches if τ_t and τ_c work in different directions, violating the monotonicity assumption, for example. Structural estimators can be expected to perform better when they are more closely linked to prevailing theoretical arguments. We do not wish to read too much into the relative performance of each of the three methods we considered in this particular analysis.

5. CONCLUSION

We began by observing that one potential reason for the disconnect between the behavioral patterns anticipated by formal models of international conflict and those identified in the historical record using statistical analysis of quantitative data is an inherent inability to control for important confounding variables. We cannot know for certain whether the theoretical models that indicate that would-be targets of aggression can often prevent war by arming actually capture enough important elements of the true data-generating process to have correctly identified the true relationship between arming and war. The important point here is that *even if* they did, we would still expect to observe empirical evidence that *appeared* to contradict the implications of such models.[56]

Specifically, we presented a simple extension of the canonical ultimatum bargaining model

[56]We illustrated this problem by focusing on military arming, though states have many other options available to them (Powell 1999, Trager 2010), and the logic of our argument should apply more broadly. We suspect it also applies to the study of other important political and economic phenomena, such as the effectiveness of attack ads or attempts to stimulate aggregate demand.

in which a Target has the option of arming at the outset. The results indicated that by arming, the Target increases the level of confidence the Challenger must have that the Target is weak before the Challenger is willing to make a demand that risks war. We then created simulated data assuming that the bargaining model perfectly described the interactions between states engaged in crisis bargaining — which we assumed most pairs of states were not. We further assumed that it is impossible to determine which states engage in crisis bargaining, although we assumed both arming and conflict are observable.

If we could restrict our analysis to only those dyads engaged in crisis bargaining, we would correctly infer that arming generally prevents war. However, we cannot restrict our analysis in this way, and for good reason. There is no way of knowing how often we observe peace simply because one side revised the status quo in their favor, knowing that they risked war in so doing, and was fortunate enough to have their gamble pay off.

Many observable indicators, such as those based on geography or the record of past tensions, might help us to identify dyads that are more likely to engage in crisis bargaining than others. Such variables *may* facilitate correct inferences, provided they serve as relatively good proxies. Perhaps our assumptions about the quality of available proxies were unduly pessimistic. If so, one might imagine that simple correctives, such as pre-processing using matching, would be perfectly capable of facilitating correct inferences. However, we caution that, for reasons discussed above, our analysis likely *underestimates* the overall problem. Moreover, we found that the bias associated with matching depends substantially on the choice of dependent variable. One would need to believe that we have greatly *overestimated* the problem in order to conclude that the inferences one would draw after matching on observables would be correct when the dependent variable represents a low level of conflict, such as a militarized interstate dispute. The degree to which we would need to have overestimated the severity of the problem before one would conclude that we are unduly critical of the promise of matching would be considerably lower with respect to analyses employing war as the dependent variable. However, given the rarity of war, it is unlikely that scholars will

find a research design that requires them to ignore large numbers of observations to be practical. We therefore conclude that scholars would have to believe we have badly overstated the severity of the problem, even though we have gone to considerable lengths to be sure to understate it, if they are to justify the use of matching when employing dependent variables that measure levels of hostility short of war when seeking to test hypotheses concerning the likelihood of war.

Supposing they are persuaded by our critique of matching, scholars might alternatively seek out new variables to instrument for potentially endogenous regressors. Our results suggest this approach might well be capable of correctly identifying the negative relationship between arming and conflict, *even* when conflict is measured using a variable that records lower levels of hostility. However, several important issues that were of little concern for our stylized Monte Carlo simulations can impede the use of instrumental variables in practice. For one, proper instruments are very difficult to identify. Even when exogeneity is plausible, if the instrument is only a weak predictor of the endogenous regressor, it can introduce considerable bias (Bound, Jaeger, and Baker 1995). Bias may also be introduced by limitations in the optimization routines of canned bivariate probit estimators in commonly used statistical software packages (Freedman and Sekhon 2010). Additionally, instruments only allow for the identification of local treatment effects, which can be an important limitation in some contexts. In sum, identifying proper instruments may bear fruit, but we stress that instrumental variables are no panacea.

More practically, one might assume a very basic structure for the strategic interactions between states and employ existing statistical estimators in a manner consistent with the logic of backwards induction. We found that a relatively simplistic application of statistical backwards induction, where we uncritically adopted a model nearly identical to that of Bas, Signorino, and Walker (2008), failed to correctly identifying a negative relationship between arming and conflict. Of course, Bas, Signorino, and Walker (2008) gave us little reason to believe that the simple model they used to illustrate the power of their approach would be

the appropriate one in all circumstances. We urge scholars to consider structural estimators that are more closely linked to prevailing theoretical arguments pertaining to their specific research question.

Ultimately, when dealing with observational data, we must always be concerned that those observations that exhibit certain characteristics are more likely to experience the outcomes we seek to explain due to factors other than the characteristics we observe. This problem need not be insurmountable, but our analysis offers little hope to those in search of an approach that requires the analyst neither to think carefully about the underlying theory nor to make any strong assumptions. The challenges to correct causal inference are difficult to overstate.

6. APPENDIX A

Remark 1. There are eight pure strategy perfect Bayesian equilibria of the model, all of which are pooling equilibria. In the interest of space, we focus here only on characterizing the conditions under which the equilibria hold. See the web appendix for a more detailed proof.

Proof. Begin with the equilibria in which the Target pools on arming.

When $w' \leq \overline{w}$ and $w'' \leq \underline{w}$, the Challenger sets $x = \underline{p} + c_2 \equiv \underline{x}$ if the Target arms and $x = \tilde{p} + c_2 \equiv \tilde{x}$ otherwise. As long as $\tilde{p} - \underline{p} \geq \kappa$, incentive compatibility is satisfied for both the weak and strong types and the equilibrium holds.

When $w' > \overline{w}$ and $w'' \leq \underline{w}$, the Challenger sets $x = \hat{p} + c_2 \equiv \hat{x}$ if the Target arms and $x = \tilde{x}$ otherwise. For the weak type of Target, incentive compatibility requires $\tilde{p} - \hat{p} \geq \kappa$. The strong type's behavior follows the equilibrium as long as $\tilde{p} - \underline{p} \geq \kappa$. If $\tilde{p} - \hat{p} \geq \kappa$, this latter condition is satisfied, and the equilibrium holds.

When $w' \leq \overline{w}$ and $w'' > \underline{w}$, the Challenger sets $x = \underline{x}$ if the Target arms and $x = \overline{p} + c_2 \equiv \overline{x}$ otherwise. Incentive compatibility requires $\overline{p} - \underline{p} \geq \kappa$ for the weak type, and $\tilde{p} - \underline{p} \geq \kappa$ for the strong type. If $\overline{p} - \underline{p} \geq \kappa$, the latter condition is satisfied, and the equilibrium holds.

When $w' > \overline{w}$ and $w'' > \underline{w}$, the Challenger sets $x = \hat{x}$ if the Target arms and $x = \overline{x}$ otherwise. Incentive compatibility requires $\overline{p} - \hat{p} \geq \kappa$ for the weak type and $\tilde{p} - \underline{p} \geq \kappa$ for the strong type. As long as both inequalities hold, the equilibrium holds.

Now consider the equilibria in which the Target pools on not arming. If we reverse the sign of each of the inequalities defining incentive compatibility above, we readily identify the conditions for the four equilibria in which neither type arms. It is straightforward to establish that sufficiently large values of κ satisfy each of these conditions. This demonstrates that there are eight pure strategy perfect Bayesian equilibria, all of which are pooling.

We now demonstrate that there cannot be any separating equilibria. There are two potential separating equilibria.

Let us first consider the case where the strong type arms and the weak does not. By Bayes' rule, the Challenger's posterior beliefs must be $w' = 0$ and $w'' = 1$. Given these beliefs, the Challenger sets $x = \underline{x}$ if the Target arms and $x = \overline{x}$ otherwise. Incentive compatibility for the weak type requires $\kappa \geq \overline{p} - \underline{p}$ and $\tilde{p} - \underline{p} \geq \kappa$ for the strong type. If $\overline{p} - \underline{p} > \tilde{p} - \underline{p}$, κ cannot simultaneously be large enough to satisfy the first condition and small enough to satisfy the second. Yet since $\overline{p}$ is the largest value of p and $\underline{p}$ is the smallest, the left side of the inequality must be strictly larger than the right side. Therefore, it must always be true that either the strong type has an incentive to deviate or the weak type does.

Next, consider the case where the weak type arms and the strong type does not. The Challenger's posterior beliefs must therefore be $w' = 1$ and $w'' = 0$, and the Challenger sets $x = \hat{x}$ if the Target arms and $x = \tilde{x}$ otherwise. Incentive compatibility for the weak type requires $\tilde{p} - \hat{p} \geq \kappa$ and $\kappa \geq \tilde{p} - \underline{p}$ for the strong type. If $\tilde{p} - \underline{p} > \tilde{p} - \hat{p}$, κ cannot simultaneously be large enough to satisfy the former and small enough to satisfy the latter. This simplifies to $\hat{p} > \underline{p}$, which must be true. Thus, this equilibrium fails, completing the proof. □

7. APPENDIX B

In this appendix, we provide more detail on how we generated the *WAR* variable. As discussed in Appendix A, the equilibria to the game are uniquely identified by two beliefs held by the Challenger: w' and w''. To represent these beliefs, we independently generate $WPRIME \sim U[0,1]$ and $WPRIMEPRIME \sim U[0,1]$. In order to create *WOVER*, which represents $\overline{w}$, and *WUNDER*, which represents $\underline{w}$, we construct *PROBOVER*, *PROBHAT*, *PROBTILDE*, and *PROBUNDER*, representing $\overline{p}$, $\hat{p}$, $\tilde{p}$, and $\underline{p}$, respectively.

For each observation where *CRISIS* equals 1:

$$
ARMEQB \begin{cases}
= 1 & \text{if } PROBOVER - PROBHAT > KAPPA, \\
 & PROBTILDE - PROBUNDER > KAPPA, \\
 & WPRIME > WOVER, \text{ and } WPRIMEPRIME > WUNDER \\
= 1 & \text{if } PROBOVER - PROBUNDER > KAPPA, \\
 & PROBTILDE - PROBUNDER > KAPPA, \\
 & WPRIME < WOVER, \text{ and } WPRIMEPRIME > WUNDER \\
= 1 & \text{if } PROBTILDE - PROBHAT > KAPPA, \\
 & PROBTILDE - PROBUNDER > KAPPA, \\
 & WPRIME > WOVER, \text{ and } WPRIMEPRIME < WUNDER \\
= 1 & \text{if } PROBTILDE - PROBUNDER > KAPPA, \\
 & WPRIME < WOVER, \text{ and } WPRIMEPRIME < WUNDER \\
= 0 & \text{Otherwise.}
\end{cases}
$$

We then generate a binary variable, *BOLD*, equal to 1 if the Challenger sets x to a value the Target accepts if and only if the Target is weak. This is true when $WPRIME > WOVER$ if $ARMEQB = 1$ and when $WPRIMEPRIME > WUNDER$ if $ARMEQB = 0$. Next, we generate a binary variable, *WEAK*, equal to 1 with probability 0.5. Finally, *WAR* is a binary variable equal to 1 if $BOLD = 1$ and $WEAK = 0$.

REFERENCES

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Bas, Muhammet Ali, Curtis S. Signorino, and Robert Walker. 2008. "Statistical Backwards Induction: A Simple Method for Estimating Recursive Strategic Models." *Political Analysis* 16(1): 21–40.

Bennett, D. Scott, and Allan C. Stam. 2004. *The Behavioral Origins of War*. Ann Arbor, MI: University of Michigan Press.

Bound, John, David Jaeger, and Regina Baker. 1995. "Problems with Instrumental Variable Estimation When the Correlation Between Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430): 443–450.

Clarke, Kevin. 2009. "Return of the Phantom Menace: Omitted Variable Bias in Political Research." *Conflict Management and Peace Science* 26(1): 46–66.

Colaresi, Michael, and William Thompson. 2005. "Alliances, Arms Buildups and Recurrent Conflict: Testing a Steps-to-War Model." *Journal of Politics* 67(2): 345–364.

Diehl, Paul, and Mark Crescenzi. 1998. "Reconfiguring the Arms Race-War Debate." *Journal of Peace Research* 35(1): 111–118.

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

Filson, Darren, and Suzanne Werner. 2002. "A Bargaining Model of War and Peace: Anticipating the Onset, Duration and Outcome of War." *American Journal of Political Science* 46(4): 819–838.

Fortna, Virginia Page. 2004. "Does Peacekeeping Keep Peace: International Intervention and the Duration of Peace After Civil War." *International Studies Quarterly* 48(2): 269–292.

Freedman, David, and Jasjeet Sekhon. 2010. "Endogeneity in Probit Response Models." *Political Analysis* 18(2): 138–150.

Gartzke, Eric. 1999. "War Is in the Error Term." *International Organization* 53(3): 567–587.

Ghosn, Faten, Glenn Palmer, and Stuart Bremer. 2004. "The MID3 Data Set, 1993-2001: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science* 21(2): 133–54.

Gibler, Douglas, Toby Rider, and Marc Hutchison. 2005. "Taking Arms Against a Sea of Troubles: Conventional Arms Races During Periods of Rivalry." *Journal of Peace Research* 42(2): 131–147.

Gilligan, Michael G., and Ernest J. Sergenti. 2008. "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference." *Quarterly Journal of Political Science* 3(2): 89–122.

Hewitt, Joseph. 2003. "Dyadic Processes and International Crises." *Journal of Conflict Resolution* 47(5): 669–692.

Imai, Kosuke, Gary King, and Olivia Lau. 2007. "Zelig: Everyone's Statistical Software.".

Imai, Kosuke, Gary King, and Olivia Lau. 2008. "Toward A Common Framework for Statistical Analysis and Development." *Journal of Computational and Graphical Statistics* 17(4): 892–913.

Jones, Daniel, Stuart Bremer, and J. David Singer. 1996. "Militarized Interstate Disputes, 1816-1992: Rationale, Coding Rules, and Empirical Patterns." *Conflict Management and Peace Science* 15(2): 163–213.

Lewis, Jeffrey, and Kenneth Schultz. 2003. "Revealing Preferences: Empirical Estimation of a Crisis Bargaining Game with Incomplete Information." *Political Analysis* 11(4): 345–367.

Morton, Rebecca. 1999. *Methods and Models*. New York, NY: Cambridge University Press.

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

Powell, Robert. 2004. "Bargaining and Learning While Fighting." *American Journal of Political Science* 48(2): 344–361.

Reed, William. 2000. "A Unified Statistical Model of Conflict Onset and Escalation." *American Journal of Political Science* 44(1): 84–93.

Sample, Susan. 1997. "Arms Races and Dispute Escalation: Resolving the Debate." *Journal of Peace Research* 34(1): 7–22.

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

Schultz, Kenneth. 2001. *Democracy and Coercive Diplomacy*. Cambridge, UK: Cambridge University Press.

Sekhon, Jasjeet. 2008. "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods." In *The Oxford Handbook of Political Methodology*, ed. Janet Box-Steffensmeier, Henry Brady, and David Collier. Oxford, UK: Oxford University Press pp. 271–300.

Sekhon, Jasjeet. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Sekhon, Jasjeet S. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R." *Journal of Statistical Software* 42(7): 1–52.

Senese, Paul, and John Vasqeuz. 2008. *The Steps to War: An Empirical Study.* Princeton, NJ: Princeton University Press.

Senese, Paul, and John Vasquez. 2005. "Assessing the Steps to War." *British Journal of Political Science* 35(4): 607–633.

Signorino, Curtis S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93(2): 279–298.

Signorino, Curtis S. 2003. "Structure and Uncertainty in Discrete Choice Models." *Political Analysis* 11(4): 316–344.

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.

Signorino, Curtis S., and Kuzey Yilmaz. 2003. "Strategic Misspecification in Regression Models." *American Journal of Political Science* 47(3): 551–566.

Slantchev, Branislav. 2003. "The Principle of Convergence in Wartime Negotiations." *American Political Science Review* 97(4): 621–632.

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

Slantchev, Branislav. 2011. *Military Threats: The Costs of Coercion and the Price of Peace.* New York, NY: Cambridge University Press.

Sovey, Allison J., and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Reader's Guide." *American Journal of Political Science* 55(1): 188–200.

Tarar, Ahmer, and Bahar Leventoğlu. 2008. "Does Private Information Lead to Delay or War in Crisis Bargaining?" *International Studies Quarterly* 52(3): 533–553.

Trager, Robert. 2010. "Diplomatic Calculus in Anarchy: How Communication Matters." *American Political Science Review* 104(2): 347–368.

Wagner, R. Harrison. 2000. "Bargaining and War." *American Journal of Political Science* 44(3): 469–484.

Wand, Jonathan. 2006. "Comparing Models of Strategic Choice: The Role of Uncertainty and Signaling." *Political Analysis* 14(1): 101–120.

Whang, Taehee. 2010. "Empirical Implications of Signaling Models: Estimation of Belief Updating in International Crisis Bargaining." *Political Analysis* 18(3): 381–402.

Xiang, Jun. 2010. "Relevance as a Latent Variable in Dyadic Analysis of Conflict." *Journal of Politics* 72(2): 484–498.

Zagare, Frank, and D. Marc Kilgour. 2000. *Perfect Deterrence*. Cambridge, UK: Cambridge University Press.

Table 1: Summary of Equilibria

	Pooling on Not Arm	Pooling on Arm
$w' \leq \overline{w}$	Challenger sets $x = \underline{x}$	Challenger sets $x = \underline{x}$
$w'' \leq \underline{w}$	$Pr(\text{War}) = 0$	$Pr(\text{War}) = 0$
$w' > \overline{w}$	Challenger sets $x = \underline{x}$	Challenger sets $x = \overline{x}$
$w'' \leq \underline{w}$	$Pr(\text{War}) = 0$	$Pr(\text{War}) = 1 - w$
$w' \leq \overline{w}$	Challenger sets $x = \overline{x}$	Challenger sets $x = \underline{x}$
$w'' < \underline{w}$	$Pr(\text{War}) = 1 - w$	$Pr(\text{War}) = 0$
$w' > \overline{w}$	Challenger sets $x = \overline{x}$	Challenger sets $x = \overline{x}$
$w'' > \underline{w}$	$Pr(\text{War}) = 1 - w$	$Pr(\text{War}) = 1 - w$

Note: w' is the Challenger's posterior belief that the Target is weak given that the Target armed and w'' is the Challenger's posterior belief that the Target is weak given that the Target did not arm.

Table 2: Summary of Results

	$\hat{\beta}$	Standard deviation	Standard error	Rejection rate $\hat{\beta}$
Baseline Logit: Experiment 1				
All Observations, DV: *MID*	0.68	0.03	0.003	1.0
All Observations, DV: *WAR*	0.27	0.05	0.005	1.0
Only *CRISIS*, DV: *MID*	−0.16	0.05	0.005	0.90
Only *CRISIS*, DV: *WAR*	−0.70	0.06	0.006	1.0
Baseline Logit: Experiment 2				
All Observations, DV: *MID*	0.73	0.03	0.003	1.0
All Observations, DV: *WAR*	0.40	0.05	0.005	1.0
Only *CRISIS*, DV: *MID*	−0.16	0.05	0.005	0.90
Only *CRISIS*, DV: *WAR*	−0.70	0.06	0.006	1.0
Matching: Experiment 1				
DV: *MID*	0.69	0.09	0.009	1.0
DV: *WAR*	−0.23	0.14	0.014	0.41
Matching: Experiment 2				
DV: *MID*	0.84	0.10	0.01	1.0
DV: *WAR*	0.08	0.16	0.016	0.05

Notes: Results based on estimation of 100 independent data sets. Results only reported for ARM. For matching models, results reported after matching on $PROX_1 = PROX_2 = 1$.

Table 3: Predicted Probabilities (Baseline Logit, Matching, SBI Models)

	$ARM = 0$	$ARM = 1$
Baseline Logit: Experiment 1		
All Observations, DV: *MID*	**0.4309**	**0.5985**
	(0.4174, 0.4445)	(0.5837, 0.6133)
All Observations, DV: *WAR*	**0.1286**	**0.1623**
	(0.1172, 0.1400)	(0.1465, 0.1781)
Only *CRISIS*, DV: *MID*	**0.8087**	**0.7830**
	(0.7987, 0.8187)	(0.7704, 0.7957)
Only *CRISIS*, DV: *WAR*	**0.2362**	**0.1331**
	(0.2254, 0.2470)	(0.1227, 0.1435)
Baseline Logit: Experiment 2		
All Observations, DV: *MID*	**0.2674**	**0.4310**
	(0.2558, 0.2790)	(0.4151, 0.4469)
All Observations, DV: *WAR*	**0.0688**	**0.0997**
	(0.0611, 0.0765)	(0.0876, 0.1117)
Only *CRISIS*, DV: *MID*	**0.8087**	**0.7830**
	(0.7987, 0.8187)	(0.7704, 0.7957)
Only *CRISIS*, DV: *WAR*	**0.2362**	**0.1331**
	(0.2254, 0.2470)	(0.1227, 0.1435)
Matching: Experiment 1		
DV: *MID*	**0.4627**	**0.6324**
	(0.4385, 0.4869)	(0.5985, 0.6663)
DV: *WAR*	**0.1292**	**0.1055**
	(0.1130, 0.1456)	(0.0839, 0.1271)
Matching: Experiment 2		
DV: *MID*	**0.2767**	**0.4701**
	(0.2548, 0.2987)	(0.4288, 0.5114)
DV: *WAR*	**0.0706**	**0.0761**
	(0.0580, 0.0831)	(0.0542, 0.0980)
Statistical Backwards Induction: Experiment 1		
All Observations, DV: *MID*	**0.4218**	**0.6180**
	(0.4175, 0.4262)	(0.6133, 0.6225)
Only *CRISIS*, DV: *MID*	**0.8082**	**0.7841**
	(0.8070, 0.8094)	(0.7827, 0.7855)
Statistical Backwards Induction: Experiment 2		
All Observations, DV: *MID*	**0.2608**	**0.4495**
	(0.2575, 0.2641)	(0.4433, 0.4546)
Only *CRISIS*, DV: *MID*	**0.8083**	**0.7844**
	(0.8070, 0.8096)	(0.7830, 0.7859)

Notes: Results based on estimation of 100 independent data sets. For matching models, results reported after matching on $PROX_1 = PROX_2 = 1$. Statistical backwards induction models that only include observations where $CRISIS = 1$ exclude $PROX_1$ and $PROX_2$ and are based on CAP_1 held at mean value of 0.5.

Table 4: Predicted Probabilities (Instrumental Variables)

	$ARM = 0$	$ARM = 1$
Experiment 1		
All Observations, DV: *MID*		
$Y_1 = 0$, $Y_2 = 0$	**0.3591**	**0.3396**
	(0.3482, 0.3700)	(0.3282, 0.3511)
$Y_1 = 0$, $Y_2 = 1$	**0.2484**	**0.2678**
	(0.2389, 0.2580)	(0.2573, 0.2785)
$Y_1 = 1$, $Y_2 = 0$	**0.1897**	**0.1769**
	(0.1814, 0.1981)	(0.1685, 0.1855)
$Y_1 = 1$, $Y_2 = 1$	**0.2029**	**0.2157**
	(0.1945, 0.2114)	(0.2065, 0.2249)
All Observations, DV: *WAR*		
$Y_1 = 0$, $Y_2 = 0$	**0.5475**	**0.5982**
	(0.5351, 0.5600)	(0.5856, 0.6107)
$Y_1 = 0$, $Y_2 = 1$	**0.0594**	**0.0086**
	(0.0540, 0.0650)	(0.0069, 0.0105)
$Y_1 = 1$, $Y_2 = 0$	**0.2377**	**0.3437**
	(0.2269, 0.2487)	(0.3312, 0.3564)
$Y_1 = 1$, $Y_2 = 1$	**0.1554**	**0.0495**
	(0.1467, 0.1644)	(0.0435, 0.0559)
Experiment 2		
All Observations, DV: *MID*		
$Y_1 = 0$, $Y_2 = 0$	**0.4893**	**0.4543**
	(0.4777, 0.5099)	(0.4420, 0.4665)
$Y_1 = 0$, $Y_2 = 1$	**0.1697**	**0.2048**
	(0.1618, 0.1778)	(0.1950, 0.2146)
$Y_1 = 1$, $Y_2 = 0$	**0.2258**	**0.2056**
	(0.2163, 0.2353)	(0.1965, 0.2149)
$Y_1 = 1$, $Y_2 = 1$	**0.1156**	**0.1361**
	(0.1096, 0.1218)	(0.1290, 0.1433)
All Observations, DV: *WAR*		
$Y_1 = 0$, $Y_2 = 0$	**0.6228**	**0.6533**
	(0.6116, 0.6360)	(0.6410, 0.6655)
$Y_1 = 0$, $Y_2 = 1$	**0.0350**	**0.0055**
	(0.0311, 0.0391)	(0.0044, 0.0069)
$Y_1 = 1$, $Y_2 = 0$	**0.2544**	**0.3150**
	(0.2433, 0.2655)	(0.3030, 0.3271)
$Y_1 = 1$, $Y_2 = 1$	**0.0868**	**0.0262**
	(0.0804, 0.0933)	(0.0224, 0.0302)

Notes: Results based on estimation of 100 independent data sets. Results based on $IV_1 = 1$ and $PROX_1 = PROX_2 = 1$.

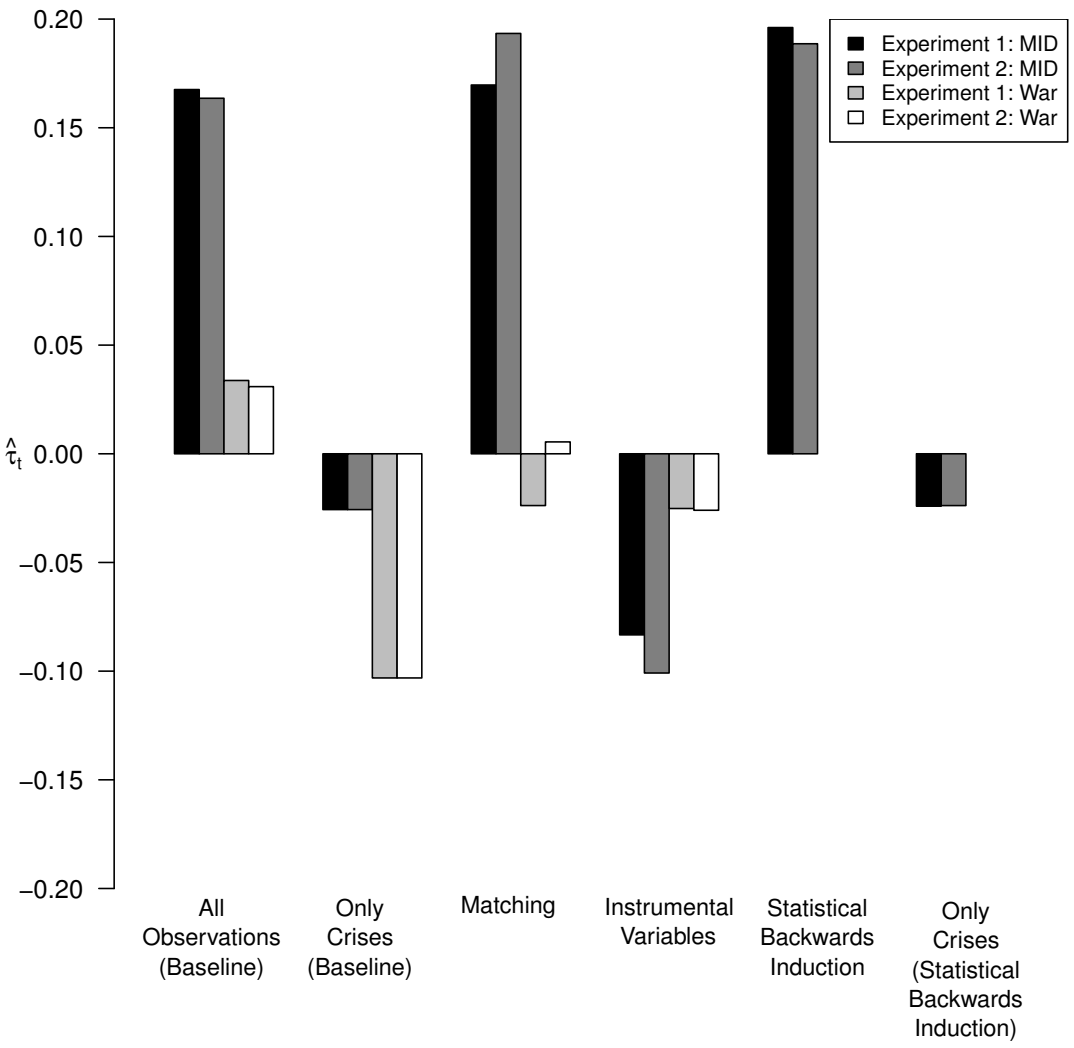

Figure 1: Comparison of Baseline Models and Three Empirical Methods