Political Science Research and Methods Vol 3, No. 3, 569–587 September 2015
© The European Political Science Association, 2015
doi:10.1017/psrm.2014.42

Challenges to Inference in the Study of Crisis Bargaining*

PHILIP ARENA AND KYLE A. JOYCE

The possibility that actors strategically condition their behavior on partially unobservable factors poses a grave challenge to causal inference, particularly if only some of the actors whose behavior we analyze are at risk of experiencing the outcome of interest. We present a 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 interactions 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 (and unsurprisingly) indicate a positive relationship between arming and war. More importantly, we then evaluate the performance of matching, instrumental variables, and statistical backwards induction. The latter two show some promise, but matching fares poorly.*

Unilaterally seizing control of vital resources, particularly those that legally belong to other states, is highly provocative. Wars have started over smaller offenses. Yet, such actions do not constitute a threat, display, or use of force. When states claim disputed resources, risking violent responses, the incidents often do not appear in prominent data sets analyzed by scholars of international conflict. For example, China recently "announced plans to start regular civilian patrols of the disputed South China Sea, a move that could further escalate tensions in the region," not long after passing a regulation "requiring foreign fishermen to seek Beijing's approval to operate in much of the South China Sea" (*Voice of America* 2014). These moves drew condemnation from neighboring states, such as Vietnam, but thus far, China's *fait accompli* has not provoked any militarized response. Similarly, India recently began building hydroelectric dams on rivers the 1960 Indus Waters Treaty allocated to Pakistan, and this has led Hafiz Saeed, who is thought to have planned the 2008 terrorist attacks in Mumbai, to accuse India of "water terrorism" and to threaten that either "water flows or blood" (*The Age* 2012). At this point in time, Pakistan has not responded militarily, appears unlikely to do so (*Reuters* 2012), and no attacks comparable with those in Mumbai have occurred; however, India could not have known for certain that a flagrant violation of the treaty would not lead to militarized violence.

It is well understood that if some units face fundamentally different risks of experiencing an outcome, failure to account for such differences can bias our estimates of the effects of other explanatory variables. If the factors that account for different risks of exposure are *observable*, the remedy is straightforward. We might control for those factors or pre-process the data with matching methods, but what if such factors are *unobservable*? We can turn to proxy variables (i.e., observable

* Philip Arena, Visiting Assistant Professor of Political Science, Department of Political Science, University of Rochester, Harkness Hall, Rochester, NY 14627 (filarena99@gmail.com). Kyle A. Joyce, Assistant Professor of Political Science, Department of Political Science, University of California, Davis One Shields Ave., 469 Kerr Hall, Davis, CA 95616 (kjoyce@ucdavis.edu). The authors thank Amber Boydstun, Daina Chiba, Sean Gailmard, Heather McKibben, Randy Siverson, Zeev Maoz, Jamie Monogan III, Bill Reed, Toby Rider, Jas Sekhon, and Curt Signorino for valuable comments. The authors also thank Matt Buttice for research assistance. This article also benefited from comments provided by the participants at EITM (2010) and attendees of the Annual Meeting of the Society for Political Methodology (2010). To view supplementary material for this article, please visit http://dx.doi.org/10.1017/psrm.2014.42

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

factors we have reason to believe correlate with the omitted variables), or, if a good instrument can be identified, we can use instrumental variables. However, we are unlikely to draw correct inferences if such fixes only partially account for differences among the observations.

There are two reasons international conflict scholars should find this prospect quite troubling. First, the most prominent explanation for war implies some states face systematically different risks of war, and that the source of this variation is largely unobservable. Bargaining models of war indicate that when challengers are unable to determine whether the target will accept or reject an offer, their actions balance the risk of war against the benefit of securing a better distributive outcome, with strictly unobservable factors separating peaceful incidents in which no state threatened, showed, or used force, yet where one side understood their actions created a non-trivial risk of war (as in the examples above), from cases in which there was never any meaningful chance of war. Second, we do not know much about how statistical techniques typically employed to address such problems perform under the conditions that international conflict scholars likely face in practice.

The broader point is that one of the most prominent approaches to explaining war suggests we cannot assume that "crises" always begin with observable actions of the sort recorded by existing data sets. There is strong theoretical reason to believe that such near-misses occur frequently enough that existing data sets may not provide an adequate basis for determining which dyads have the potential for conflict. This poses a profound challenge to our ability to correctly diagnose the relationship between the actions states take and the likelihood they subsequently find themselves at war. Any behavior states undertake during crises in order to prevent escalation, but would rarely undertake otherwise, will be positively correlated with the incidence of conflict.

The concern that some zeros fundamentally differ from other zeros is not new. This very issue has played a prominent role in previous debates about the determinants of conflict. Scholars have investigated the degree to which sampling on politically relevant dyads can alleviate the concern (Lemke and Reed 2001). Another solution is to treat the problem as omitted variable bias, where the usual solution is to include a laundry list of observable control variables, despite the problems with this approach (Clarke 2009). We show the full extent of the problem is not yet appreciated, as many of the standard fixes fail to solve the problem under conditions we believe to be relatively conservative below.

We illustrate our argument in the context of one important and prominent example: military arming. Slantchev (2011) argues that states can often deter aggression through military preparations for war. Yet, empirical research often either fails to find a consistent relationship or finds that arming is positively associated with war (Gibler, Rider and Hutchison 2005; Rider, Findley and Diehl 2011). We argue that extant research is insufficient to allow us to determine whether such findings contradict the theoretical expectation because even if our theoretical models were *exactly* right about how arming affects crisis bargaining, the typical research design could fail to identify this relationship. In order to show this result, we first present a crisis bargaining model in which arming typically reduces the likelihood of war. We then create a simulated data set generated directly from our theoretical model and assess the prospects for identifying the true negative relationship between arming and conflict through generalized linear models, matching methods, instrumental variables, and statistical backwards induction (SBI).

We demonstrate, unsurprisingly, that using a naïve research design tends to lead us to the wrong conclusion.[1] The problem often does not go away after matching on observables, particularly if we use a dependent variable that measures low-level conflicts, as virtually all quantitative studies of

[1] Notably, the inclusion of variables that proxy for the true confounder may fail to ensure correct inferences. This is true even if the proxy variable captures a considerable amount of the variation in the omitted variable.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

international conflict do.[2] However, our results indicate that using instrumental variables can often uncover the true negative relationship, regardless of how conflict is measured. Finally, a simplistic application of SBI, one that follows closely the example provided by Bas, Signorino and Walker (2008), fails to identify the true negative relationship.

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 assume the true relationship is strictly monotonic, that all observable variables are measured without 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 (DGP) but fully characterizes the path to war. We hope few scholars would make such optimistic assumptions when they work with observational data.

Our results suggest the following 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 probably not the best answer to the problems we identify. It may be worthwhile to search out valid instruments—though we are cognizant of well-known limitations to their use (Bound, Jaeger and Baker 1995). Finally, we strongly caution against employing the simplest possible approach to structural estimation without considering whether the underlying theoretical model is appropriate, though this does not detract from the value of truly *theoretically informed* structural estimators—which we do not analyze. Overall, these results indicate that purely inductive inquiry, or attempts to "let the data speak," may often produce incorrect inferences. In particular, scholars should be wary about drawing inferences about the impact of behaviors chosen by actors whose risk of exposure to the outcome of interest varies. Whether such concerns apply to a given topic of inquiry is, we believe, best determined with the assistance of well-developed theory.

MODEL

The game begins with Nature selecting the Target's military capability m_2. Nature sets $m_2 = \underline{m}_2$ with probability w, indicating the Target is relatively weak, and $m_2 = \overline{m}_2$ with probability $1-w$, indicating the Target is relatively strong, where $0 < \underline{m}_2 < \overline{m}_2$. The Target then chooses whether or not to arm. If the Target arms, it incurs cost $k > 0$, and its total military capabilities increase by $\alpha > 0$.

After observing the Target's decision, the Challenger demands $x \in [0,1]$, which the Target either accepts or rejects. If the Target accepts, the Challenger receives x whereas the Target receives either $1-x$ or $1-x-k$, depending on whether the Target armed. If the Target rejects, a war occurs, which the Challenger wins with probability p. We assume the side that wins keeps the entire value of the good in dispute. Moreover, we assume each side incurs some loss of utility owing to the costs of war, denoted $c_1 \in (0,1]$ and $c_2 \in (0,1]$ for the Challenger and Target, respectively. 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 $1-p-c_2-k$, if the Target armed).

The precise value of p is a function of m_1, m_2, and, if the Target armed, α, where $m_1 > 0$ is the Challenger's military capability. If the Target is weak and did not arm, $p = \frac{m_1}{m_1+\underline{m}_2} \equiv \overline{p}$. If the Target is weak and armed, $p = \frac{m_1}{m_1+\underline{m}_2+\alpha} \equiv \hat{p}$. If the Target is strong and did not arm,

[2] 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.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

$p = \frac{m_1}{m_1+\overline{m}_2} \equiv \tilde{p}$. If the Target is strong and armed, $p = \frac{m_1}{m_1+\overline{m}_2+\alpha} \equiv \underline{p}$. 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.[3]

EQUILIBRIA

As we assume the Challenger is uninformed about the Target's type, we solve the model for perfect Bayesian equilibria. Thus, the Challenger's beliefs must incorporate the information revealed, if any, by the Target's initial decision over whether to arm, and the Challenger's strategies must be optimal, both on- and off-the-equilibrium path, given their posterior beliefs. There are eight such equilibria in pure strategies, all of which are pooling equilibria.[4] There are four equilibria in which both the strong and weak type Target pool on not arming and four in which both the strong and weak type Target pool on arming.[5]

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 there are only pooling equilibria, the Target's decision to arm does not influence the Challenger's beliefs. Yet, arming still plays an essential role. Critically, the Challenger is willing to risk war under more restrictive conditions after witnessing the Target arm than the Challenger is when the Target does not arm.

Formally, when the Target arms, the Challenger sets $x = \overline{x}_A$ when $w' > \frac{c_1+c_2}{\tilde{p}-p+c_1+c_2} \equiv \overline{w}$ and $x = \underline{x}_A$ when $w' \leq \overline{w}$, where w' denotes the Challenger's posterior belief the Target is weak given the Target armed. If the Target does not arm, the Challenger sets $x = \overline{x}_N$ when $w'' > \frac{c_1+c_2}{\tilde{p}-\tilde{p}+c_1+c_2} \equiv \underline{w}$, and sets $x = \underline{x}_N$ when $w'' \leq \underline{w}$, where w'' denotes the Challenger's posterior belief the Target did not arm. Note that in any given pooling equilibrium, either w' or w'' will simply equal w.[6]

In addition, note the relatively small $(\underline{x}_A)$ and large $(\overline{x}_A)$ demands the Challenger makes after observing the Target arm are distinct from the relatively small $(\underline{x}_N)$ and large $(\overline{x}_N)$ demands the Challenger makes if the Target does not arm. This difference accounts for the difference between $\underline{w}$ and $\overline{w}$. 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 type. The key result 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 did not arm.[7]

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 there are conditions under which the probability of war is unaffected by the Target's decision. Specifically, when $w' \leq \overline{w}$ and $w'' \leq \underline{w}$ or when $w' > \overline{w}$ and $w'' > \underline{w}$, the Target's decision has no impact on the probability of war. In the former case, war never occurs; in the latter case, the Challenger risks war regardless of the Target's decision.

In contrast, when $w' \leq \overline{w}$ and $w'' > \underline{w}$, the Challenger risks war if and only if the Target did not arm. In such cases, we would conclude that arming strictly prevents war. However, when

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

[4] This implies that when arming prevents war, it does not do so by revealing information, or acting as a costly signal, as it does in Slantchev (2011).

[5] Proofs are in web appendix.

[6] One of the Challenger's posterior beliefs must always mirror the Challenger's prior belief, given there are only pooling equilibria.

[7] *Proof.* If $\overline{p} - \tilde{p} > \hat{p} - \underline{p}$, it must be true that $\overline{w} > \underline{w}$, as $\underline{w}$ has a larger denominator. Substituting the values of p, we obtain $\frac{m_1}{m_1+\overline{m}_2} - \frac{m_1}{m_1+\overline{m}_2} > \frac{m_1}{m_1+\overline{m}_2+\alpha} - \frac{m_1}{m_1+\overline{m}_2+\alpha}$. As the right-hand side of the inequality is identical to the left-hand side, except 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.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

TABLE 1 *Summary of Equilibria*

	Pooling on Not Armed	Pooling on Armed
$w' \leq \overline{w}$	Challenger sets $x = \underline{x}_N$	Challenger sets $x = \underline{x}_A$
$w'' \leq \underline{w}$	$P_r(\text{War}) = 0$	$P_r(\text{War}) = 0$
$w' > \overline{w}$	Challenger sets $x = \underline{x}_N$	Challenger sets $x = \bar{x}_A$
$w'' \leq \underline{w}$	$P_r(\text{War}) = 0$	$P_r(\text{War}) = 1 - w$
$w' \leq \overline{w}$	Challenger sets $x = \bar{x}_N$	Challenger sets $x = \underline{x}_A$
$w'' > \underline{w}$	$P_r(\text{War}) = 1 - w$	$P_r(\text{War}) = 0$
$w' > \overline{w}$	Challenger sets $x = \bar{x}_N$	Challenger sets $x = \bar{x}_A$
$w'' > \underline{w}$	$P_r(\text{War}) = 1 - w$	$P_r(\text{War}) = 1 - w$

Note: w' is the Challenger's posterior belief the Target is weak given the Target is armed and w'' is the Challenger's posterior belief the Target is weak given the Target did not arm. Cell entries report the demand the Challenger makes in equilibrium when the Target either pools on not arming (second column) or pools on arming (third column).

$w' > \overline{w}$ and $w'' \leq \underline{w}$, the Challenger risks war if and only if the Target does not arm. In such cases, arming *promotes* war.

However, unless we make some unusual assumptions about the beliefs held by the Challenger with respect to events that never occur in equilibrium, we can nonetheless conclude that arming *more often than not* prevents war. For example, suppose the Challenger's off-the-equilibrium path belief matches the Challenger's prior belief. As $\underline{w} < \overline{w}$, 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 they are distributed similarly to the Challenger's prior beliefs in order to conclude the conditions under which arming strictly reduces the probability of war will be met more often than the conditions under which arming promotes war. As we discuss below, this is precisely what we assume in our simulated data.

If and *only* if both types of Target arm, and the Challenger's prior belief the Target is weak is sufficiently large, then the Challenger is willing to risk war with a relatively large demand, but the Challenger would believe the Target is less likely to be weak if the Target did not arm, then we would conclude that arming promotes war. It is unclear why the Challenger would be expected to hold such a belief as the stronger Target does not face a greater incentive to deviate from the equilibrium strategy of arming.

Nonetheless, in our simulated data, we allow the Challenger's out-of-equilibrium belief to take on any value with equal probability, and still conclude arming reduces the probability of war. Therefore, there are *some* observations for which the equilibrium probability of war is non-zero, but had the Target deviated from the equilibrium strategy and chosen not to arm, the probability of war would be zero. For this reason, we stress the true relationship between arming and war in our simulated data is that, *on average*, arming reduces the probability of war. There are more cases where this is true than where it is not.

Scholars who are not persuaded by our treatment of out-of-equilibrium beliefs should note, irrespective of whether such assumptions would be warranted if our goal was to use the model to generate observable implications that were then tested with observational data, given the way we set up our Monte Carlo simulations, we must nonetheless conclude we are being misled by statistical models that fail to identify that, *within our simulated data*, arming generally reduces the probability of war.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

IMPLICATIONS

The equilibria discussed above have several important properties useful for our Monte Carlo simulations, where we treat the model as the complete DGP. We do so not because we believe this to be plausible, but rather because it illustrates the problem more clearly.[8]

First, the model generates equilibria in which war occurs despite the Target choosing to arm. As we evaluate a probabilistic relationship between arming and conflict, the absence of such equilibria could introduce the zero-likelihood problem. 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, although 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. For any given set of parameter values, we can uniquely identify each player's strategies and the resulting probability of war.

It is useful to express the implications of the model more precisely using the potential outcomes framework. Let Y_{i1} be a binary indicator for whether observation i experiences conflict given the Target armed (is treated) and Y_{i0} be a binary indicator for whether i experiences conflict given 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 what *would* have happened if the state did not arm. Similarly, if a state does not arm, we cannot observe what *would* have happened had it armed. As we cannot observe counterfactual states of the world, two of the four quantities of interest are unobservable, and we cannot directly assess the causal effect of the treatment on any given observation.

What 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 are owing to the fact one set of states armed and one did not. However, this only facilitates correct inferences if we correct for all other relevant differences between those states that armed and those that did not. This is no small caveat, given we have theoretical reason to believe we cannot hope to perfectly distinguish empirically between states engaged in crisis bargaining and those that are not, nor between instances where states unilaterally alter the status quo to minimize the risk of war versus those where states believed they pushed their opponent to the brink, yet were fortunate to discover the other side was willing to allow their *fait accompli* to stand.

We focus on the average treatment effect for the treated (ATT), rather than the average treatment effect for the control (ATC). The ATT (τ_t) is $\tau|(T = 1) = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1)$ and ATC (τ_c) is $\tau|(T = 0) = E(Y_{i1}|T_i = 0) - E(Y_{i0}|T_i = 0)$, where $T_i = 1$ indicates that i is in the treatment regime (armed) and $T_i = 0$ that i is in the control regime (did not arm). That is, we focus on whether observations where the Target armed would have experienced a lower probability of conflict if the Target did not arm. This is related to, but distinct from, the question of whether the probability of war among those observations where the Target did not arm would have been higher had the Target armed. We focus on the former because, although the effects are substantively similar in both cases, the ATT should be larger and the failure to properly identify it would be more striking than the failure to properly identify the ATC.[9]

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

[9] There will be more cases where the probability of war is precisely 0 in the control group than in the treatment group because treatment status is correlated with whether the states are engaged in crisis bargaining. This will attenuate the relationship between arming and war in the control group.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

Yet, as mentioned above, there may be reasons to suspect that it is difficult to determine which states are engaged in crisis bargaining. As crisis bargaining is a predictor of both war and arming, we have reason to doubt that extant empirical analysis has identified the true relationship between arming and war.

We are not claiming to know that arming prevents war. It is entirely possible that the theoretical models that suggest it does are over-simplified. Our argument is simply that *even if it were true* that arming prevents war, we might falsely conclude that it does not—and this is a much harder problem to solve than is commonly appreciated, even when using techniques designed to solve such problems.

Table 1 shows that when $w' \leq \overline{w}$ and $w'' \leq \underline{w}$ or when $w' > \overline{w}$ and $w'' > \underline{w}$, the probability of war does not depend on the Target's decision to arm. In the language of potential outcomes, that means there are cases for which $\tau_t = E(Y_{i1}) - E(Y_{i0}) = 0$. The same will be true for all observations where the states did not engage in crisis bargaining, as the probability of war in these cases is 0 by assumption, regardless of treatment status.

There also exist cases where $E(Y_{i1}) - E(Y_{i0}) > 0$ as well as cases where $E(Y_{i1}) - E(Y_{i0}) < 0$.[10] That is, 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 web appendix), we can be confident the τ_t and τ_c 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.

SIMULATED DATA

We generate the military capabilities for State 1 and State 2 for 100,000 dyads.[11] 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 $\epsilon \sim N(1,0.1)$. In 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 $ALPHA \sim N(0.5,0.05)$ for α, $KAPPA \sim N(0.5,0.005)$ for k, and $COST_i \sim N(0.1,0.01)$ for $i \in \{1,2\}$ for c_i.[12]

BARG is a binary variable indicating whether the states in the dyad engaged in crisis bargaining, which is true for ≈10 percent of the observations (randomly selected). We assume crisis bargaining is perfectly described by the model.[13] Ultimately, $WAR = 1$ in 20 percent of the observations where $BARG = 1$. Given we assume the path to war is perfectly captured by our theoretical model, when $BARG = 0$, war is never observed.

If State 2 arms in order to deter one potential aggressor, their arming decision will be reflected in all dyads containing State 2, as we cannot readily determine which state prompted an increase in State 2's military capabilities. We create a binary variable *ARM*, which equals 1 when $BARG = 1$ and the formal model indicates State 2 would arm, and equals 1 when $BARG = 0$ with probability 0.25 if $KAPPA \leq \overline{KAPPA}$, and with probability 0.1 if $KAPPA > \overline{KAPPA}$. Overall, $ARM = 1$ in 41 percent of the observations when $BARG = 1$ and 18 percent when $BARG = 0$.

IV_1 and IV_2 are instruments for *ARM*. $IV_1 = 1$ with probability 0.75 if $KAPPA \leq 0.05$ and 0 otherwise. IV_2 is constructed identically, except the probability is 0.9. $IV_1 = 1$ ≈34 percent of

[10] These cases exist in both the treatment and control groups.

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

[12] The values we set for the parameters in the simulated data are representative of other values we examined. Overall, the results are qualitatively similar, only varying in how pronounced the effects were.

[13] In order to determine equilibrium behavior, we also generated variables representing the Challenger's beliefs (see web appendix).

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

the time when $ARM = 0$, and 53 percent of the time when $ARM = 1$. For IV_2, these figures are 40 and 64 percent, respectively.

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

A few characteristics of *MID* 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 16 percent of our observations when $MID = 1$. In contrast, <4 percent of the *MIDs* in the data used by Bennett and Stam (2004) are wars, for example. Our results will therefore *underestimate* the problems 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, $BARG = 0$ for a substantial number of observations where our simulated *MID* variable equals 1 ($\approx$36 percent). The MID data were designed to minimize the possibility of such cases. "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" (Jones, Bremer and Singer 1996, 168). Thus, our *MID* variable may correlate with *BARG* to a lesser extent than is the case in the actual data. It is important to note this also ensures we are *underestimating* the problems associated with testing hypotheses about the causes of war using the MID data. As *MID* and *BARG* become less distinct, we become more likely to estimate a positive relationship between *ARM* and *MID*, given that *BARG* predicts *ARM*.

We also create a variable 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 $BARG = 1$, $MID = 1$, and $WAR = 0$, and with probability 0.075 if $BARG = 0$, $MID = 1$, and $WAR = 0$. *RECIP* is coded as missing if $MID = 0$. Approximately 43 percent of our simulated *MIDs* involve reciprocated hostility.

Although we assume *BARG* is unobservable, we generate two binary 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. 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 set $PROX_1 = 1$ with probability 0.5 if $BARG = 1$, and probability 0.1 if $BARG = 0$, whereas $PROX_2 = 1$ with probability 0.3 if $BARG = 1$, and probability 0.1 if $BARG = 0$. We refer to the data set containing these variables as Experiment 1.

Next, we generate a simulated data set identical to the first one, except $PROX_1 = 1$ with probability 0.4 if $BARG = 1$, and probability 0.15 if $BARG = 0$, whereas $PROX_2 = 1$ with probability 0.2 if $BARG = 1$, and probability 0.1 if $BARG = 0$. We refer to this data set as Experiment 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 *BARG*, which is not possible in practice.[14] For our

[14] *BARG* 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, *BARG* is a clear confounder, as it is directly related to the likelihood of war, while the cost of arming is not. Also, note that conditioning on *BARG* is tantamount to matching. However, when we evaluate matching methods, we only match on observable variables as this is all that can be done in practice.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

TABLE 2 *Summary of Results (Logit, Matching)*

	$\hat{\beta}$	SD	SE	Rejection rate $\hat{\beta}$
Logit: Experiment 1				
All observations: *MID*	0.68	0.02	0.002	1.0
All observations: *WAR*	0.27	0.06	0.006	1.0
Only *BARG*: *MID*	–0.16	0.05	0.005	0.90
Only *BARG*: *WAR*	–0.70	0.06	0.006	1.0
Logit: Experiment 2				
All observations: *MID*	0.73	0.02	0.002	1.0
All observations: *WAR*	0.40	0.06	0.006	1.0
Only *BARG*: *MID*	–0.16	0.05	0.005	0.90
Only *BARG*: *WAR*	–0.70	0.06	0.006	1.0
Matching: Experiment 1				
MID	0.69	0.10	0.001	1.0
WAR	–0.24	0.15	0.001	0.42
Matching: Experiment 2				
MID	0.87	0.10	0.001	1.0
WAR	0.05	0.20	0.002	0.07

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

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 Experiments 1 and 2. We initially include all dyads, as is common in applied research. Next, we match on *BARG*. For these models, we do not include our proxy variables as they are no longer necessary to isolate the systematic difference between the treatment and control groups.

Table 2 summarizes the results from 100 independently created data sets, each of which 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.[15] We also report the standard error $\left(\frac{s}{\sqrt{100}}\right)$ and the rejection rate of $\hat{\beta}$ based on H_o: $\beta = 0$ versus H_a: $\beta \neq 0$ using a two-tailed test and significance level of $p < 0.05$.[16]

The top panel of Table 2 contains the results. When the analysis includes all dyads, we *always* estimate a positive relationship between arming and conflict, irrespective of whether conflict is measured using *MID* or *WAR*, for both Experiments 1 and 2. We also always reject the null hypothesis when analyzing all dyads.

In contrast, when we condition on *BARG*, we *always* observe the negative relationship anticipated by the formal model, regardless of the dependent variable and for both Experiments 1 and 2.[17] In addition, we *always* reject the null hypothesis when we condition on *BARG* and *WAR* is the dependent variable, and nearly always when *MID* is the dependent variable. This clearly illustrates the problem is our inability to determine who is engaged in crisis bargaining.

Note 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

[15] 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. 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 $BARG = 1$ (results in web appendix).

[16] 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.

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

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

TABLE 3 *Predicted Probabilities (Logit, Matching, and SBI Models)*

	ARM = 0	*ARM* = 1
Logit: Experiment 1		
All observations: *MID*	0.433 (0.419, 0.447)	0.599 (0.585, 0.615)
All observations: *WAR*	0.129 (0.119, 0.141)	0.163 (0.147, 0.179)
Only *BARG*: *MID*	0.810 (0.800, 0.820)	0.784 (0.771, 0.797)
Only *BARG*: *WAR*	0.238 (0.227, 0.249)	0.134 (0.123, 0.144)
Logit: Experiment 2		
All observations: *MID*	0.269 (0.257, 0.280)	0.432 (0.416, 0.448)
All observations: *WAR*	0.069 (0.062, 0.077)	0.099 (0.088, 0.112)
Only *BARG*: *MID*	0.810 (0.800, 0.820)	0.784 (0.771, 0.797)
Only *BARG*: *WAR*	0.238 (0.227, 0.249)	0.134 (0.123, 0.144)
Matching: Experiment 1		
MID	0.464 (0.440, 0.488)	0.633 (0.599, 0.667)
WAR	0.130 (0.114, 0.146)	0.106 (0.084, 0.127)
Matching: Experiment 2		
MID	0.276 (0.254, 0.298)	0.476 (0.434, 0.517)
WAR	0.071 (0.059, 0.084)	0.075 (0.054, 0.097)
SBI: Experiment 1		
All Observations: *MID*	0.423 (0.419, 0.427)	0.620 (0.616, 0.625)
Only *BARG*: *MID*	0.810 (0.809, 0.811)	0.784 (0.783, 0.786)
SBI: Experiment 2		
All observations: *MID*	0.261 (0.257, 0.265)	0.447 (0.441, 0.452)
Only *BARG*: *MID*	0.810 (0.809, 0.811)	0.785 (0.783, 0.786)

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

effect. Table 3 shows the predicted probabilities and 95 percent confidence intervals when $PROX_1 = PROX_2 = 1$, which corresponds to the case where states are most likely to arm, from which we can estimate $\hat{\tau}_t$.[18] When we include all dyads, the change in the size of the effect is larger for *MID* compared with *WAR*. However, the change in the size of the effect when we condition on *BARG* is larger for *WAR* compared with *MID*. *ARM* decreases the probability of *WAR* by 0.10 among those dyads engaged in crisis bargaining. However, among the same dyads, *ARM* only decreases the probability of *MID* by 0.03.

Despite including control variables ($PROX_1$ and $PROX_2$) we *always* estimate a positive $\hat{\tau}_t$ when analyzing all dyads. Once we condition on *BARG* we correctly identify a negative $\hat{\tau}_t$ using a binary logit. The problem is we cannot condition on this variable in practice.

MATCHING

We now discuss the results after matching on $PROX_1 = PROX_2 = 1$.[19] We implicitly employed exact matching (using *BARG*), and correctly identified a negative $\hat{\tau}_t$.[20] However, matching does not

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

[19] Analogous models were estimated with $PROX_1 = PROX_2 = 0$, which corresponds to cases where *ARM* is least likely to equal 1, and allows us to estimate $\hat{\tau}_c$. 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 presented below is that $\hat{\tau}_t$ is always positive. The $\hat{\tau}_t$ is also smaller than the results presented below when *MID* is the dependent variable (results available in web appendix).

[20] Given that *WAR* never occurs when *BARG* equals 0, had we attempted to estimate $\hat{\tau}_c$, we would find that the outcome variable does not vary.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

reliably uncover treatment effects when confounders are unobservable, as this violates the assumption of conditional independence (Sekhon 2008).[21] We know we have violated this assumption. In practice, researchers cannot know whether they have done so. All they can do is demonstrate their matched data set is balanced with respect to *observable* covariates and hope for the best.

We focus on exact matching, which is the preferred approach when feasible.[22] Exact matching isolates much, but not all, of the difference between treatment and control groups. In our simulated data, *BARG* is 2.8 times as likely to equal 1 when $ARM = 1$ before matching, versus 1.46 after matching in Experiment 1 and 1.94 in Experiment 2.[23]

The bottom panel of Table 2 summarizes the results.[24] In the baseline analysis, the primary difference between the models where the dependent variable was *MID* and those where it was *WAR* was the magnitude of the effect of *ARM*. Here, in both Experiments 1 and 2, we estimate a positive relationship between *ARM* and *MID* and *always* reject the null hypothesis. When *WAR* is 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. Thus, if we select a more appropriate dependent variable, we might at least avoid drawing the opposite conclusion, though we would fail to uncover the pacifying effect of arming.

The middle panel of Table 3 shows the predicted probabilities and 95 percent confidence intervals when $ARM = 0$ and $ARM = 1$. *ARM* increases the probability of *MID* by 0.17 and 0.20 in Experiments 1 and 2, respectively, and increases the probability of *WAR* by 0.004 in Experiment 2, but *decreases* the probability of *WAR* by 0.02 in Experiment 1. In the baseline analysis *ARM* decreased the probability of *WAR* by 0.10, five 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 generalized linear models. Such practice does not justify the use of "all else equal" language, so common in applied research. Our proxy variables are capable, under some conditions, of leading us to the correct inference when we used them for matching, but did not do so when included as controls in the baseline analysis.

Note, we have addressed neither the Stable Unit Treatment Value Assumption (SUTVA), nor whether our conditioning variables were determined before treatment status; here neither is a concern; however, in practice, they very well might be.[25] Thus, we stress that whereas matching sometimes enabled us to correctly identify a negative $\hat{\tau}_t$, it did not always, and many of the assumptions we made would be untenable in practice.

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

[22] 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 (Sekhon 2009). However, as 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.

[23] Before matching, $BARG = 1$ in 21 percent of observations when $ARM = 1$ and 7 percent when $ARM = 0$. After matching, the corresponding percentages are 80 and 55 percent, respectively, in Experiment 1 and 58 and 30 percent, respectively, in Experiment 2.

[24] The sign on $\hat{\beta}$ tells us the direction of $\hat{\tau}_t$ as the only effect of *ARM* is captured by $\hat{\beta}$.

[25] SUTVA posits that Y_{i0}, $Y_{i1} \perp\!\!\!\perp T_j \forall\, i \neq j$. There is little reason to believe the Target's ability to deter the Challenger by arming depends on how many other states in the system arm. It is not clear the same could be said of other variables of interest to 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.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

INSTRUMENTAL VARIABLES

Unlike matching methods, instrumental variables can enable researchers to uncover the true causal effect even when there are unobserved confounders.[26] 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 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).[27]

For Experiments 1 and 2, we estimate four bivariate probit models.[28] We model *ARM* as a function of the binary instrument and the proxy variables, and conflict as a function of *ARM* and the proxy variables. The log-likelihood function for each model is

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

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 the error term for the second stage, ν the error term for the first stage, and $\Phi_b(\cdot,\cdot,\rho_{\varepsilon\nu})$ the bivariate normal cumulative distribution function with correlation coefficient $\rho_{\epsilon\nu}$.

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

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

Equation 2 defines $\hat{\tau}_t$ as the difference in the expected probability of conflict given the Target armed and the expected probability of conflict given the Target did not arm divided by the expected probability the Target arms. Thus, $\hat{\tau}_t$ can be identified readily from the predicted probabilities, which are presented in Table 4. We calculate $\hat{\tau}_t$ when $IV_1 = PROX_1 = PROX_2 = 1$, thereby maximizing the probability of treatment.[29]

For Experiment 1, when Y_i is given by *MID*, we calculate $\hat{\tau}_t \approx -0.08$.[30] We calculate $\hat{\tau}_t \approx -0.03$ for *WAR*. For Experiment 2, $\hat{\tau}_t$ is –0.10 and –0.03, respectively. Thus, $\hat{\tau}_t$ is consistently negative, although surprisingly, the estimated effect is larger for *MID* than *WAR*.[31]

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

[27] 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 $D_{1i} - D_{0i} \leq 0 \,\forall\, i$.

[28] 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 $Y_1 = 1$. See also Xiang (2010), whose approach is similar to ours. We used Zelig to estimate our bivariate probit models (Imai, King and Lau 2008).

[29] We also calculated predicted probabilities where $PROX_1 = PROX_2 = 0$ (results in web appendix).

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

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

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

TABLE 4 *Predicted Probabilities (Instrumental Variables)*

	$ARM = 0$	$ARM = 1$
Experiment 1		
All observations: *MID*		
$Y_1 = 0, Y_2 = 0$	0.357 (0.347, 0.368)	0.338 (0.327, 0.349)
$Y_1 = 0, Y_2 = 1$	0.250 (0.240, 0.259)	0.269 (0.259, 0.280)
$Y_1 = 1, Y_2 = 0$	0.189 (0.181, 0.198)	0.177 (0.168, 0.185)
$Y_1 = 1, Y_2 = 1$	0.204 (0.195, 0.212)	0.216 (0.207, 0.226)
All observations: *WAR*		
$Y_1 = 0, Y_2 = 0$	0.546 (0.534, 0.599)	0.597 (0.585, 0.610)
$Y_1 = 0, Y_2 = 1$	0.060 (0.055, 0.065)	0.009 (0.007, 0.011)
$Y_1 = 1, Y_2 = 0$	0.238 (0.227, 0.249)	0.344 (0.331, 0.356)
$Y_1 = 1, Y_2 = 1$	0.156 (0.147, 0.165)	0.050 (0.044, 0.056)
Experiment 2		
All observations: *MID*		
$Y_1 = 0, Y_2 = 0$	0.488 (0.476, 0.499)	0.453 (0.440, 0.465)
$Y_1 = 0, Y_2 = 1$	0.170 (0.162, 0.179)	0.205 (0.196, 0.215)
$Y_1 = 1, Y_2 = 0$	0.226 (0.217, 0.236)	0.206 (0.197, 0.215)
$Y_1 = 1, Y_2 = 1$	0.116 (0.110, 0.122)	0.136 (0.129, 0.143)
All observations: *WAR*		
$Y_1 = 0, Y_2 = 0$	0.622 (0.610, 0.635)	0.652 (0.640, 0.664)
$Y_1 = 0, Y_2 = 1$	0.035 (0.031, 0.039)	0.006 (0.004, 0.007)
$Y_1 = 1, Y_2 = 0$	0.255 (0.244, 0.266)	0.316 (0.304, 0.328)
$Y_1 = 1, Y_2 = 1$	0.087 (0.081, 0.094)	0.026 (0.023, 0.030)

Note: results based on 100 independent data sets. Results based on $IV_1 = 1$ and $PROX_1 = PROX_2 = 1$.

These results are only somewhat encouraging. We implicitly made some strong assumptions and we did not discuss local treatment effects nor how we might identify an instrument for arming.[32] These concerns are difficult to ignore when analyzing real data.[33]

SBI

Finally, we assess the prospects for identifying the negative relationship between arming and conflict using structural estimation, specifically 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 (see Signorino 1999; Signorino 2003; Signorino and Yilmaz 2003).

We assume a similar game structure to that presented in 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 reciprocate, where reciprocation produces a conflict.

[32] The causal effects estimated using instruments are called local treatment effects because they depend on the instrument. The estimated effect pertains only to observations who would enter the treatment regime if they also receive the instrument but would remain in the 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.

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

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

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.[34] It is difficult to determine the importance of this difference. For example, it is possible 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 each player's payoff for conflict depends on whether the Target armed previously. This is very likely an important difference. We stress we do not intend the analysis 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. Our goal is to demonstrate the pitfalls associated with conducting empirical analysis divorced from prevailing theories and that this problem may not be solved by more sophisticated methods. In short, we believe 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 that if it were possible to observe *BARG*, one could uncover the correct inference even using the game structure that underlies our application of SBI.[35]

We specify the utilities at each terminal node as follows.[36] As Bas, Signorino and Walker (2008) point out, it is necessary to constrain one of the utilities to 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, and the Target's utility for conceding. The remaining utilities (the payoffs for each player in the event the Target reciprocates, and the Challenger's utility for the Target conceding) consist of three parts: a coefficient, a set of regressors, and an error term. 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 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 $X_{tR3} = PROX_2$.[37]

We require the player's choices to be probabilistic. Therefore, we assume the Target reciprocates if $X_{tR}\beta_{tR} + \epsilon_t > 0$ and the Challenger initiates an *MID* if $(1 - p_r)(X_{cC}\beta_{cC}) + (p_r)(X_{cR}\beta_{cR} + \epsilon_c) > 0$, where p_r is the probability 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

[34] 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.

[35] We also 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.

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

[37] 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.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

cannot know the true value of ϵ_c and ϵ_t, whereas 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 sub-optimal or non-Nash play, or private information held by actors who behave optimally given their information.[38] 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 (Bas, Signorino and Walker 2008, 27).

The effect of *ARM* is ambiguous. By construction, *ARM* has two important effects. Most obviously, setting $X_{cR1} = (1 - ARM)$ ensures the Challenger's utility for reciprocation is decreasing in *ARM*. However, *ARM* is also in X_{tR}. Provided $\beta_{tR1} > 0$, this effect increases p_r. Whether an increase in p_r increases or decreases the probability that $MID = 1$ is unclear, as $X_{cC}\beta_{cC}$ may be either larger or smaller than $X_{cR}\beta_{cR}$.[39]

We estimate $\hat{p}_c$ and $\hat{p}_r$ as follows. We first estimate a logit with *RECIP* as the dependent variable. We evaluate the probability of reciprocation using X_{tR}. The predicted probabilities generated from this model produce $\hat{p}_r$. Next, we construct $Z_{cC} = (1-\hat{p}_r)X_{cC}$ and $Z_{cR} = \hat{p}_r X_{cR}$. We then estimate a second logit with *MID* as the dependent variable and Z_{cC} and Z_{cR} as the independent variables. This ensures 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.[40] As the regressors in the second logit are predicted probabilities, we calculate bootstrapped standard errors in order to avoid biased estimates for the standard errors of $\hat{\beta}_{cC}$ and $\hat{\beta}_{cR}$.[41]

The bottom panel of Table 3 shows the predicted probabilities of *MID* and 95 percent confidence intervals when $PROX_1 = PROX_2 = 1$, which corresponds to the case where states are most likely to arm, for Experiments 1 and 2.[42] We held CAP_1 at its mean value of 0.5.[43] In Experiment 1, ARM increases the probability of a *MID* ($\hat{p}_c$) by 0.20. In Experiment 2, *ARM* increases the probability of *MID* by 0.19. These results indicate that SBI, using a specification similar to that in Bas, Signorino and Walker (2008), is unlikely to correctly identify the relationship between *ARM* and conflict. However, we wish to be clear about the source of the problem, which is not a fundamental flaw in the logic underlying SBI.

To illustrate this, we follow the same procedure as above, with two changes. First, we only focus on observations where $BARG = 1$. Second, because we have little reason to expect $PROX_1$ or $PROX_2$ to capture important variation in this subset of observations, we drop X_{tR2} and X_{tR3} from X_{tR}. These results show that arming decreases the probability of a *MID* by 0.03 in both Experiments 1 and 2. This reinforces our claim that the problem is that *BARG* is unobservable rather than a fundamental shortcoming of SBI.

We also stress that we focus 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

[38] The distinction is immaterial here because the Target's behavior does not depend on 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 as 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. In the current formulation, the Target need only know their own payoffs, and the Challenger must be uncertain about the Target's strategy.

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

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

[41] We used 100 non-parametric bootstrap iterations.

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

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

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

adaptation of Bas, Signorino and Walker (2008) does, rather than to assess the performance of structural estimators that are theoretically motivated, which we have every reason to believe would perform much better.

SUMMARY OF RESULTS

Throughout this article, we focussed on $\hat{\tau}_t$, the ATT. Put differently, we focussed 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 $\hat{\tau}_t$ for our baseline analysis and each of the three statistical methods discussed above.

Overall, the results from the bivariate probit models with instrumental variables performed best, consistently estimating a negative $\hat{\tau}_t$, regardless of the dependent variable or the quality of the proxy variables. Our simplistic approach to SBI fared worst of all. Matching on observables uncovered a negative effect for Experiment 1 when the dependent variable was *WAR*.

There are important lessons to be learned from these results. First, if *BARG* 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. Second, the method that performed best was the only one that explicitly modeled the Target's decision to arm. This suggests that

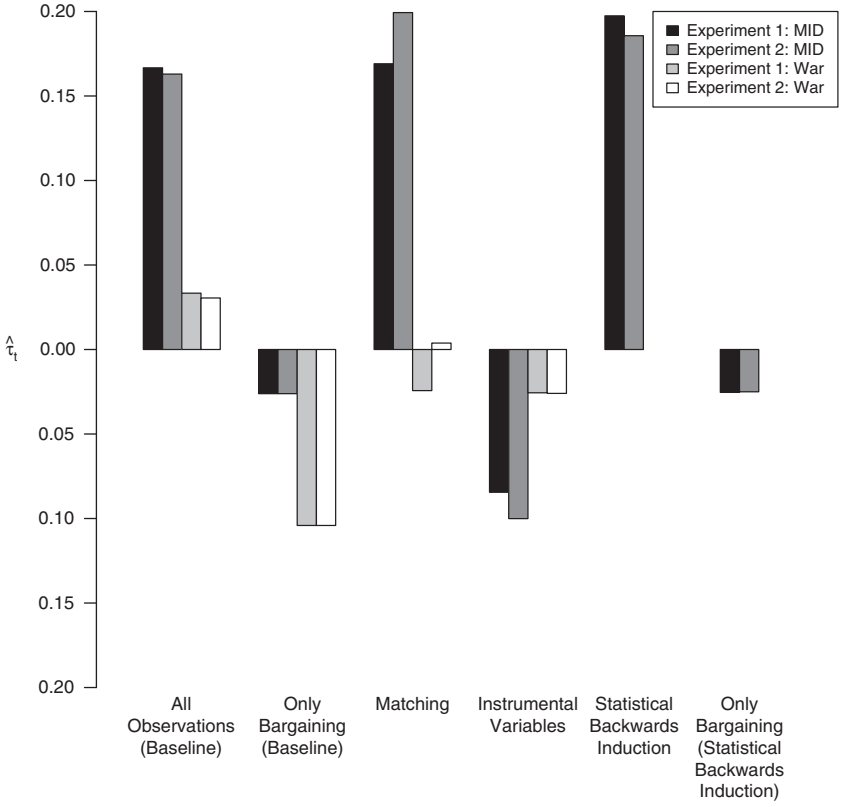


Fig. 1. Comparison of baseline models and three empirical methods

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

scholars who hope to simply "let the data speak" without thinking carefully about how the data were generated are likely to draw incorrect inferences.

However, things may be different in practice. Instrumental variables may not outperform the other approaches if $\hat{\tau}_t$ and $\hat{\tau}_c$ work in different directions, for example, as this would violate the monotonicity assumption. In addition, structural estimators that are closely linked to strong theoretical arguments should perform better.

CONCLUSION

We began by observing that states often engage in behaviors that generate a real risk of war yet which do not involve any overtly hostile acts. This complicates any attempt to separate observations where the probability of war is essentially zero from those where it is non-trivial. Absent an ability to differentiate between the two cases, we are likely to observe a positive correlation between the onset of conflict and any behavior that in fact reduces the probability of conflict. To be sure, we cannot know whether arming actually prevents war. But *if* it did, we might still observe the opposite pattern.

Though we illustrated this problem by focussing on military arming, the logic of our argument should apply more broadly. Any attempt to evaluate other deterrent strategies will face the same problem, as might the study of other political and economic phenomena, such as the effectiveness of attack adds or attempts to stimulate aggregate economic demand. If the viability of challengers to incumbents varies in ways that are difficult for political scientists to account for using existing data, and incumbents prefer not to go negative unless their prospects are dire enough that desperate measures become attractive, then the average vote share of those incumbents who go negative will be lower than the average vote share of those who do not *even if the expected vote share of those who go negative is always higher than would be true for that same candidate if they did not*. Similarly, if certain fiscal or monetary policies are only attempted when governments have reason to believe their economies are headed into rough waters, and their beliefs are based in part on information not available to social scientists, then we would expect any analysis of the relationship between stimulative policies and future growth to be biased in favor of concluding that such policies are ineffective.

To illustrate this point, we presented a bargaining model 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 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.

Of course, if we could restrict our analysis to those dyads engaged in crisis bargaining, we would correctly infer that arming generally prevents war. However, we cannot do so. What game theorists refer to as "bargaining" often takes the form of one state issuing a *fait accompli*—revising the status quo unilaterally in the hopes that such a revision will not draw a hostile response. There is no way of knowing how often peace is obtained simply because such gambles paid off.

To be sure, observable indicators might help us identify dyads that are more likely to engage in crisis bargaining than others. Variables based on geography or past tensions *may* facilitate correct inferences, provided they serve as good proxies. However, we believe our analysis likely *underestimates* the severity of this problem.

Another way of addressing the problem is to identify a proper subsample through the use of matching methods. However, if we employ measures of conflict short of war as our dependent

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

variable, such as *MIDs*, then matching on observables is unlikely to facilitate appropriate inferences. The degree to which we would need to have overestimated the severity of the problem before our criticism of matching would lose its bite is much lower for analyses employing war as the dependent variable. However, given the rarity of war, it is unlikely scholars will find a research design that requires them to ignore large numbers of observations to be practical. We therefore conclude that matching offers little hope to scholars who have reason to fear they face the problem we have identified.

Assuming 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 has considerably more promise than matching, *even* when conflict is measured using a variable that records lower levels of hostility.[44]

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. The model presented in Bas, Signorino and Walker (2008) can be readily adapted to study a range of problems. However, it might not always be the best model to use, and when it is not, analysis based on it will likely arrive at incorrect conclusions. Of course, Bas, Signorino and Walker (2008) gave us little reason to believe the simple model they used to illustrate the power of their approach would appropriate in all circumstances. We urge scholars to consult the theoretical literature pertaining to their specific research question and employ structural estimators that best reflect the arguments found therein.

Ultimately, when dealing with observational data, we must always be concerned that some observations are more likely to experience the outcomes we seek to explain owing to factors other than the ones we can measure. 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. If we are to understand the effects of political behavior, we must think carefully about why actors behave as they do.

REFERENCES

The Age. 2012. 'India's Dam Plans Anger Pakistan, Symbolise Global Water Woes'. *The Age*, 19 May.

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):443–50.

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

[44] However, several important issues 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) note that canned bivariate probit routines often fail to identify global maxima. In addition, 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 instrumental variables are no panacea.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press

Freedman, David, and Jasjeet Sekhon. 2010. 'Endogeneity in Probit Response Models'. *Political Analysis* 18(2):138–50.

Gibler, Douglas, Toby J. 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):131–47.

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.

Lemke, Douglas, and William Reed. 2001. 'The Relevance of Politically Relevant Dyads'. *Journal of Conflict Resolution* 45(1):126–44.

Lewis, Jeffrey, and Kenneth Schultz. 2003. 'Revealing Preferences: Empirical Estimation of a Crisis Bargaining Game with Incomplete Information'. *Political Analysis* 11(4):345–67.

Reuters. 2012. 'Thirsty South Asia's River Rifts Threaten "Water Wars"'. *Reuters*, 25 July.

Rider, Toby J., Michael G. Findley, and Paul F. Diehl. 2011. 'Just Part of the Game? Arms Races, Rivalry, and War'. *Journal of Peace Research* 48(1):85–100.

Sekhon, Jasjeet S. 2008. 'The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods'. In Janet M. Box-Steffensmeier, Henry E. Brady and David Collier (eds), *The Oxford Handbook of Political Methodology*. 271–99. New York, NY: Oxford University Press.

——. 2009. 'Opiates for the Matches: Matching Methods for Causal Inference'. *Annual Review of Political Science* 12:487–508.

——. 2011. 'Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R'. *Journal of Statistical Software* 42(7):1–52.

Signorino, Curtis S. 1999. 'Strategic Interaction and the Statistical Analysis of International Conflict'. *American Political Science Review* 93(2):279–98.

——. 2003. 'Structure and Uncertainty in Discrete Choice Models'. *Political Analysis* 11(4):316–44.

Signorino, Curtis S., and Kuzey Yilmaz. 2003. 'Strategic Misspecification in Regression Models'. *American Journal of Political Science* 47(3):551–66.

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.

Voice of America. 2014. 'China to Start Regular Patrols From Disputed South China Sea Island'. *Voice of America*, 21 January.

Wand, Jonathan. 2006. 'Comparing Models of Strategic Choice: The Role of Uncertainty and Signaling'. *Political Analysis* 14(1):101–20.

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):484–98.

https://doi.org/10.1017/psrm.2014.42 Published online by Cambridge University Press