Limitations to the Direct Testing of Extensive Form Crisis Bargaining Games

comments welcome, version 0.1

Jeffrey B. Lewis    Kenneth A. Schultz
Department of Political Science, UCLA
January 2002

Abstract

This paper investigates the possibility of embedding extensive-form crisis bargaining games directly into empirical tests of these interactions. Given a game structure and data on the relative frequency of the possible outcomes, what can we infer about the distribution of payoffs and, by extension, the effect of covariates (e.g., military capabilities, regime types, etc.) on those payoffs? We show that even with a very simple game structure, such inferences are plagued by under-identification and are highly sensitive to the assumed information structure. We present maximum likelihood estimators for inferring the payoffs of a basic three-move conflict game from the observed distribution of outcomes. Separate estimators are developed for quantal response (QRE), complete information Nash (NE), and perfect Bayesian equilibrium (PBE) solutions to the game. We develop two main results. First, the distribution of payoffs implied by a given set of outcome data strongly depends upon an untestable assumption about the structure of information. As a result, two analysts with the same model and the same data but different beliefs about the information structure can arrive at very different estimates of the payoffs — and have no way to determine empirically which is better. Second, regardless of the solution concept employed, non-innocuous identifying restrictions must be made that limit the generality of the results and make them difficult to interpret substantively. Researchers interested in testing strategic choice models of international conflict should focus on testing their models’ comparative-static predictions, while keeping in mind the limitations we identify.
1 Introduction

A large and growing literature in international relations seeks to test models of strategic choice, especially those pertaining to interstate bargaining and conflict (e.g., Bueno de Mesquita and Lalman 1992; Fearon 1994a; Smith 1996, 1999; Signorino 1999; Sartori 1998; Schultz 1999, 2001). Most of the work in this area has focused on testing the comparative-static predictions of game-theoretic models. If a defender becomes more powerful, is immediate deterrence more or less likely to fail (Fearon 1994a)? If a challenging state becomes democratic is the target of its threat more or less likely to resist (Schultz 1999; Partell and Palmer 1999)? If a state backs down in one crisis how does this affect the likelihood that it will get its way in a future crisis (Fearon 1994a; Sartori 1998)? The typical testing strategy is to find monotonic relationships between parameters in the model and observable outcomes, to find measurable indicators that serve as proxies for the parameters, and then use standard logit or probit models to see whether these indicators have the predicted effect on the outcome probability.

Recently, though, there has been increased interest in moving beyond comparative-static analysis to test statistical models that are derived more directly from the structure and payoffs of an extensive-form game (esp., Signorino 1999; see also Smith 1999). The rationale for doing so is appealing. In strategic settings, the relationship between variables of interest and observable outcomes may be more complicated than standard linear models permit us to capture. At a minimum, the functional forms of these relationships may look very different from those imposed by logit or probit models. At worst, the implied relationships may be non-monotonic, making them awkward to capture in off-the-shelf linear models. It is hoped that, by embedding the game structure directly into the empirical model, the latter can more accurately capture the rich and complex relationships that emerge in strategic settings.

In this paper, we examine the feasibility of this line of research. The problem we consider is as follows. Assume that the researcher has a game structure that models a common interaction among states—e.g., a crisis bargaining game. The researcher also has empirical data reporting the relative frequency of the realized outcomes in a given population, as well as perhaps information describing the characteristics of the states involved in each interaction, international conditions at the time, etc. From these data, and assuming that the game structure is correct, what inferences can
the researcher make about the distribution of terminal node payoffs? What inferences can he make about the effect of theoretically interesting covariates (e.g., the distribution of power, alliances, regime type, etc.) on those payoffs?¹

To address these questions, we develop maximum likelihood estimators for inferring the payoffs of a three-move crisis bargaining from the observed distribution of possible outcomes. Separate estimators are developed for three different information structures, each associated with a different equilibrium solution concept: complete information/Nash equilibrium, complete information/quantal response equilibrium, and incomplete information/perfect Bayesian equilibrium. In treating this last version, we move beyond current work in this area, which has so far considered only games of complete information (Signorino 1999; Smith 1999). An examination of the issues that arise in testing incomplete information models would seem to be warranted given the growing consensus among IR scholars that most of the interesting behavior in international crises is driven by asymmetric information and strategic communication (esp., Morrow 1989; Fearon 1994a, 1995).

The answers we find are not very encouraging. In particular, we develop two main results. First, the distribution of payoffs implied by a given set of outcome data strongly depends upon a non-testable assumption about the structure of information. Depending upon whether one assumes complete or incomplete information, a change in the relative frequency of a given outcome can lead to opposite conclusions about changes in the payoffs. This would not present much of a problem if the data permitted us to discriminate between the different information structures; unfortunately, they do not: all three estimators perform equally well on any possible data set. As a result, two analysts with the same model and the same data but different beliefs about the structure of information can arrive at very different estimates of the payoffs — and have no way to determine which set is better. This suggests that, at a minimum, researchers need to exercise caution in determining which results are or are not robust to alternative specifications. Second, regardless of the solution concept employed, non-innocuous identifying restrictions must be made. In other words, unless we arbitrarily fix the value of several parameters, a unique set of payoffs cannot be inferred from

¹It should be pointed out that independent variables might influence outcomes in ways other than through their effects on the payoffs. For example, Sartori (1998) argues that past crisis behavior can serve as a coordinating device that allows states to switch from one equilibrium to another. Schultz (1998) models the effect of democracy by assuming that democratic states have a second actor—an active opposition party—that is not present in nondemocracies.
the outcome data. Normalizations are common in many standard models, such as the probit, but we show that the restrictions necessary in this context impose a loss of generality. As a result, our estimates of the payoffs can hinge on untestable assumptions that make it difficult, if not impossible, to interpret the estimated payoffs. For example, an arbitrary assumption about how much one state values the stakes of the crisis can influence our estimate of the other state’s payoff from war.

Thus, even in very simple settings in which the game and information structure are fully specified, making inferences about payoffs from data on the outcomes requires strong and untestable identifying assumptions. This problem is exacerbated if the information structure is not known, in which case assuming the wrong information structure can lead to entirely misleading conclusions about payoffs and the effects of covariates on those payoffs. We conclude that researchers interested in testing strategic choice models of international conflict should focus on testing their models’ comparative-static predictions, while keeping in mind the limitations we identify.

This paper proceeds as follows. Section 2 presents the simple bargaining game that we use as the basis of the analysis. It also discusses the three information structures that we explore and then solves the game using the appropriate equilibrium concept for each version. Section 3 then discusses the issues of estimation and identification that arise in trying to infer payoff distributions from the observed outcomes of the game. We first consider a quasi-experimental world in which the game is played a large number of times by states that are drawn from the same distribution each time—that is, the expected value of each payoff is the same across iterations. We then ask: given the relative frequency of outcomes, can we recover these mean values? After establishing some of the basic issues of estimation and identification that arise in this simple context. Section 5 then concludes with some thoughts about how researchers should proceed in the light of our findings.

2 The Model

The analysis in this paper is based on the simple, three-step bargaining game shown. While more complicated games can and have been used to study international conflict, this game is the simplest possible model that allows for the possibility of signaling. We believe that the problems we identify in estimating an empirical model based on this simple game only become more intractable with more complicated structures. The extensive form is depicted in Figure 1. Two states, A and B,
have a dispute over some contested good. We assume, without loss of generality, that the good belongs to $B$ in the status quo. The game begins with a decision by $A$ whether or not to challenge $B$ for the good. A challenge is assumed to involve an explicit threat to use military force in the event that $B$ does not hand over the good. If $A$ chooses not to make a challenge, the status quo ($SQ$) prevails. If $A$ does make a challenge, then $B$ must decide whether or not to resist the demand. If $B$ does not resist, it concedes the good to $A$ ($CD$), and the crisis ends peacefully. If $B$ does resist, then $A$ must decide whether or not to fight. If $A$ chooses not to fight, then it backs down from its threat ($BD$), and the good remains in $B$’s possession; otherwise, $A$ stands firm ($SF$), and the two states fight a war for possession of the good.

We let $V_A$ and $V_B$ denote the value that states $A$ and $B$ place on the good, respectively, and normalize the payoff from not having the good to zero. Thus, in the status quo, $A$ gets a payoff of zero, and $B$ gets a payoff of $V_B$. If $B$ concedes the good, then $A$ gets a payoff of $V_A$ and $B$ gets zero.\footnote{In the theoretical model, there is no loss in generality in setting both the value of $SQ$ to $A$ and the value of $CD$ to $B$ equal to zero; however, whether or not we are willing to assume that these outcomes really have identical payoffs will affect the interpretation of the empirical results.} In the event that $A$ backs down, $B$ keeps the good and, following Fearon (1994), we assume
that $A$ incurs some audience cost for having been caught in a bluff.$^3$ We let $a$ denote $A$’s payoff in this event. In general, it makes sense to assume that $a < 0$, so that that audience cost is in fact a cost; however, we will not impose this restriction. Finally, in the event of war, the states receive some payoff that is a function of their probabilities of victory, the costs of war, the value of the good, and their risk propensities (e.g., Morrow 1985). For simplicity, we let $W_A$ and $W_B$ denote the expected value of war to $A$ and $B$, respectively.

We consider three different information structures, each of which calls for a different equilibrium solution concept:

### 2.0.1 Complete Information/Nash Equilibrium (NE)

In this version of the game, we assume that all of the payoffs are common knowledge to the participants but unknown to the analyst. Because the game is played with complete information, states will arrive at the unique subgame perfect equilibrium dictated by the payoff configuration. Simple backward induction reveals the payoff orderings that are necessary and sufficient for each outcome to be realized in equilibrium:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Ordering</th>
</tr>
</thead>
<tbody>
<tr>
<td>War ($SF$)</td>
<td>$W_A &gt; a$, $W_A &gt; 0$, $W_B &gt; 0$</td>
</tr>
<tr>
<td>$A$ Backs Down ($BD$)</td>
<td>$a &gt; 0$, $W_A &lt; a$</td>
</tr>
<tr>
<td>$B$ Concedes ($CD$)</td>
<td>$W_A &gt; a$, $W_B &lt; 0$</td>
</tr>
<tr>
<td>Status Quo ($SQ$)</td>
<td>$W_A &lt; a &lt; 0$</td>
</tr>
</tbody>
</table>

Notice that the values the states attach to the good, $V_A$ and $V_B$, do not influence the equilibrium outcome in this version of the game, as long as they are both greater than zero—that is, as long as the good really is a good. This will be important later on, because it means that knowing the outcome of the game tells us nothing about the value of the good relative to the other payoffs.

---

$^3$For our purposes, it does not matter whether the audience costs are imposed by domestic audiences (Fearon 1994) or by foreign audiences (Sartori 1998).
2.1 Agent Error/Quantal Response Equilibrium (QRE)

In this version of the game, the payoffs are again common knowledge to all participants; however, states are assumed to make errors as they evaluate their choices. At each node, states make what amounts to a probabilistic choice in which the action with the higher expected utility is more likely, but not certain, to be chosen (see McKelvey & Palfrey 1995 or Signorino 1999). The strategy in solving for such a “quantal response equilibrium (QRE)” is as follows. Assume that the actor faces a choice between two actions, $a_1$ and $a_2$, and that the expected utility from $a_1$ is $U_1$ and the expected utility from $a_2$ is $U_2$. The actor is assumed to make some error in evaluating this choice, an error that is captured by adding a disturbance term to the expected value of each branch. Thus, rather than comparing $U_1$ and $U_2$, the actor compares

$$U_1^* = U_1 + \epsilon_1$$

and

$$U_2^* = U_2 + \epsilon_2.$$ 

where $\epsilon_1$ and $\epsilon_2$ are independent, normally distributed random variables with mean zero and variance $\sigma^2$. The probability that the actor chooses $a_1$ is then equal to

$$Pr(U_1^* > U_2^*)$$

or

$$\Phi \left( \frac{U_1 - U_2}{\sigma \sqrt{2}} \right)$$

where $\Phi$ is the normal cumulative distribution function. We can readily apply this solution concept to the present model. At its final node, state $A$ faces a choice between $W_A$ and $a$. The probability of fighting is

$$q_2 \equiv \Phi \left( \frac{W_A - a}{\sigma \sqrt{2}} \right)$$

State $B$ then faces a choice between zero and a lottery worth

$$q_2 W_B + (1 - q_2) V_B$$
Again perturbing each value by a normally distributed error, state $B$ will resist with probability

$$p \equiv \Phi \left( \frac{q_2 W_B + (1 - q_2) V_B}{\sigma \sqrt{2}} \right)$$

Finally, state $A$ faces a choice between the status quo and a lottery worth

$$p[q_2 W_A + (1 - q_2) a] + (1 - p)V_A$$

The probability of making the challenge is then

$$q_1 = \Phi \left( \frac{p[q_2 W_A + (1 - q_2) a] + (1 - p)V_A}{\sigma \sqrt{2}} \right)$$

Notice that each probability can be easily rendered as function of the payoffs by substituting (1) into (2) and then (2) into (3). Moreover, the probability of each observable outcome can be determined by multiplying the various choice probabilities together. Thus,

- $\Pr(SQ) = 1 - q_1$
- $\Pr(CD) = q_1(1 - p)$
- $\Pr(BD) = q_1p(1 - q_2)$
- $\Pr(SF) = q_1pq_2$

### 2.2 Incomplete Information/Perfect Bayesian Equilibrium (PBE)

In this version of the game, each state’s expected value for war is assumed to be private information. Nature selects $W_A$ and $W_B$ according to some probability distribution over the real numbers, and the realized values are known only by $A$ and $B$, respectively. The probability distributions from which these values were drawn are assumed to be common knowledge. We solve this game for a perfect Bayes Nash equilibrium in which all the strategies are sequentially rational given the actors’ beliefs, and beliefs are calculated from the equilibrium strategies according to Bayes’ rule, whenever possible. A complete derivation of the equilibrium is given in Appendix A.

The basic form of the equilibrium is as follows. A's strategy is described by a pair of cutpoints in the continuum of possible types. The first cutpoint, which is exactly equal to the audience cost
term, \(a\), separates those that will fight at the final node \((W_A > a)\) from those that will not \((W_A < a)\). The second cutpoint, which we label \(b\), separates those types that make a challenge \((W_A > b)\) and those that do not \((W_A < b)\). It is often the case that \(b < a\), in which case the continuum of types is divided into three behavioral types: those that make a challenge and fight if resisted \((W_A > a)\), those that make a challenge and back down if resisted \((a > W_A > b)\), and those that accept the status quo \((W_A < b)\). There are, however, conditions under which \(b > a\), in which case all types that make the challenge also fight if resisted (and some that would fight do not make the challenge). There are also conditions under which \(b\) is effectively negative infinity, meaning that all types make the challenge in equilibrium. The conditions under which these different patterns hold are delineated in the appendix.

State B’s response in the event of a challenge is also described by a cutpoint in the continuum of types. We let \(c\) denote the critical value of \(W_B\) such that \(B\) resists only if its value for war is above this threshold. Given these strategies, we can again determine the probabilities of each terminal node:

\[
\begin{align*}
Pr(SQ) &= Pr(W_A \leq b) \\
Pr(CD) &= Pr(W_A > b) Pr(W_B \leq c) \\
Pr(BD) &= Pr(a \geq W_A > B) Pr(W_B > c) \\
Pr(SF) &= Pr(W_A > \min[a, b]) Pr(W_B > c)
\end{align*}
\]

### 3 Issues of Estimation and Identification

We now turn to the feasibility of estimating the payoffs of the game from observational data on the frequency of each outcome. Ideally, we would like to estimate each payoff \(\theta \in \{V_A, V_B, W_A, W_B, a\}\), under the assumption that each can be written as a linear combination of a vector of covariates and a disturbance term that capture random variation from one observation to the next. That is, we assume that each payoff for each observed play of the game \(i\) can written as

\[
\theta_i = X_{i\theta} \beta_\theta + \epsilon_{i\theta},
\]

where \(X_{i\theta}\) represents a matrix of covariates of \(\theta\), \(\beta_\theta\) measures the effect of each covariate on \(\theta\), and \(\epsilon_{i\theta}\) are the random shocks. It would be ideal if consistent estimates of the \(\beta_s\) could be
made. The reality, as we will see, is that estimation of such effects is fraught with problems of identification. Moreover, the substantive conclusions about the effects of covariates on payoffs are highly contingent on the assumed information structure and equilibrium concept.

In order to demonstrate these problems in a straightforward way, we first focus on a simpler problem with an idealized data generating process. In this idealized world, each payoff for observation \( i \) can written as a constant plus a disturbance term, or

\[
\theta_i = \bar{\theta} + \epsilon_{i\bar{\theta}}.
\]

We assume that the random shocks (\( \epsilon s \)) are normally distributed with mean zero and variance \( \sigma^2 \) and independent across observations and payoffs. This process is analogous to the situation where the analyst observes the same game played repeatedly with the only random variation from observation to observation. In this case, the object of the estimation is to infer the average payoffs, \((V_A, V_B, W_A, W_B, \bar{a})\) and the variance of the shocks \( \sigma^2 \) from observations on the outcomes of plays of the game.

Immediately, however, we must scale back even this more limited ambition due to a degrees of freedom problem. The game has four possible outcomes, so the data consists of four observed frequencies: how often did the game end at \( SQ \), how often at \( CD \), how often at \( BD \), and how often at \( SF \). Because these frequencies must sum to one, the probability of any one outcome is completely determined by the probabilities of the other three. Hence, there are really only three independent observations. On the other hand, there are 6 different parameters that we need to estimate: the 5 payoff means and the variance of the disturbance terms, \( \sigma^2 \).\(^4\) Clearly, additional restrictions are needed before we can even begin.

A natural first restriction is to fix the value of the good to one \((V_B = V_A = 1)\). This normalization effectively determines the utility scale for each state. That is, for both countries keeping the good or taking it without a fight is worth 1, while not having it or giving it up without a fight is worth 0. All of the other payoffs in the game must then be considered in relation to this normalization. At one level, this is not too much of a problem. Because we assume the payoffs to the game

---

\(^4\)The list would be even longer had we not already made several simplifying assumptions: we have fixed at zero the value of the status quo to \( A \) and the value of concession to \( B \), and we have assumed that all \( \epsilon_{i\theta} \) have the same variance.
are von Neuman-Morganstern utilities, they are only defined up to a linear transformation. Fixing two of the payoffs of the game for each actor chooses a particular transformation and effectively nails down the utility scales. Only if we have to make additional restrictions — and we will — is there a loss of generality. At the same time, there are costs to this normalization. We will see later that our inability to empirically estimate $V_A$ and $V_B$ can hamper our ability to assess the effect of covariates on other payoffs. In addition, this normalization precludes the direct comparison of payoffs across states. Unless one is willing to assume that both states place the same value on the good and that the value of $SQ$ to $A$ and $CD$ to $B$ are really identical (e.g., zero), then the scales on which we measure the two states’ payoffs are different. In theoretical models, an inability to make interpersonal comparisons of utility is rarely a problem. From an empirical standpoint, however, it limits the hypotheses we can test. For example, we cannot assess whether regime type has a larger effect on the challenger’s value for war than the target’s.

The observability of the random shocks is a main point of differentiation among the three approaches presented in the previous section. With Complete Information/NE, all of the shocks in given realization of the game are assumed to be known to both $A$ and $B$. Conversely, with Agent Error/QRE the shocks represent a “perceptual disturbance” of the true payoffs, which are assumed to be the $\theta$s, and the actors behave as if new shocks are drawn at each decision node. Finally, in the Incomplete information/PBE version, the shocks associated with $W_A$ and $W_B$, $\epsilon_{W_k}$ and $\epsilon_{W_B}$, are private information for $A$ and $B$ respectively and the shock to $a$ is common knowledge.

The general likelihood function for the data is

$$L = \prod_{P \in \{SF, CD, BD, SQ\}} \Pr(P|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma)^{n_P}$$

where $n_P$ is the number times each outcome $P$ occurred. The corresponding log likelihood can be

---

5 The only loss of generality that arises from setting $V_A = V_B = 1$ is that we rule out the possibility that the good is actually not a “good” by imposing the restriction that having the good is better than not.

6 Technically, the QRE assumes a somewhat different error structure. As noted above, the QRE assumes independent shocks to the expected value of each alternative at each choice node. This is very similar, but not identical to a model in which new shocks are drawn for each payoff at each choice node. The difference is that in QRE the variance of the random component added to each expected utility is $\sigma^2$, whereas if each payoff that comprised each expected utility were shocked independently the variance of the shock to the expected utility would be $\sigma^2 \sum_j p_j^2$ where $p_j$ is the probability of the occurrence of each outcome $j = 1, \ldots, J$ included in the expectation.
written as,

$$\ln L = N \sum_P f(P) \ln \Pr(P|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma)$$  \hspace{1cm} (1)$$

where $f(P)$ is the frequency with which outcome $P$ occurs in the data and $N$ is the total number of observations. Maximizing this likelihood is nearly identical to minimizing

$$\sum_P (f(P) - \Pr(P|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma))^2.$$  \hspace{1cm} (2)$$

Intuitively, the objective is to pick values of the payoffs that, as closely as possible, reproduce the observed frequency distribution.

Each of the approaches developed in Section 2 implies a different set of probabilities of over the set outcomes for a given set of payoffs. We now describe the probability distribution of the outcomes given the payoffs for each of the three approaches. For each approach, we show that values of the payoffs exist such that the objective in (2) is made equal to zero, and, therefore, the payoffs exactly predict the observed distribution of outcomes. Thus, all three of the models are “saturated” and the data cannot be used to test hypotheses about which (if any) of the approaches was the one that generated the data.

**Complete Information/NE**

In this model, we assume that all of the payoffs are known to both of the actors. In particular, the values of each $\varepsilon$ and the average payoffs are common knowledge to the players. In this case, the actors have knowledge beyond what the observer can gain by repeated observation. While the players know the full measure of each payoff, the analyst can learn only the average payoffs. Given the equilibrium definitions described in Section 2, the probability of each outcome conditional on the average payoffs is:

$$\Pr(SF|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma) = \Phi_2 \left( \frac{\bar{W}_A}{\sigma}, \sqrt{2}(\bar{W}_A - \bar{a})/\sigma, \sqrt{2}/\sigma \right) \Phi \left( \frac{\bar{W}_B}{\sigma} \right)$$

$$\Pr(BD|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma) = \Phi_2 \left( \sqrt{2}(\bar{a} - \bar{W}_A)/\sigma, \bar{a}/\sigma, \sqrt{2}/\sigma \right)$$

$$\Pr(CD|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma) = \Phi \left( -\frac{\bar{W}_B}{\sigma} \right) \Phi \left( \frac{\sqrt{2}(W_A - a)}{\sigma} \right)$$

In “saturated” models like the ones presented here, these two objectives are identical and met with $f(P) = \Pr(P|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma)$ for $P \in \{SF, BD, CD, SQ\}$. 

---

7In “saturated” models like the ones presented here, these two objectives are identical and met with $f(P) = \Pr(P|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma)$ for $P \in \{SF, BD, CD, SQ\}$. 

---
where \( \Phi_2 \) is the standard bivariate normal cumulative distribution function and \( \Phi \) is the standard univariate normal cumulative distribution function. By definition, \( \Pr(SQ) = 1 - \Pr(SF) - \Pr(BD) - \Pr(CD) \). Thus, we can focus on the probability over three of the four payoffs with no loss of information.

Let \( T_p \) be a mapping from a set of average payoffs and residual variance to the probability of observing each outcome,

\[
T_p : (\bar{W}_A, \bar{W}_B, \bar{a}, \sigma) \rightarrow (\Pr(SF), \Pr(BD), \Pr(CD)).
\]

Note that \( T_p \) maps a system with three equations (the outcome probabilities) and four unknowns. Generally, such systems have a continuum of solutions. This is no exception. Note that \( \sigma \) appears in the denominator of every term in the equations describing the outcome probabilities. Thus, the same probabilities can always be achieved by multiplying \( \sigma \) by some (positive) constant and then multiplying \( \bar{W}_A, \bar{W}_B, \) and \( \bar{a} \) by the same constant. So, we cannot simultaneously estimate \( \sigma \) and the payoffs by this method.

To allow estimation of the payoffs, we fix \( \sigma = 1 \). This normalization is not innocuous as it is, for example, in probit regression models. Since we have already nailed down units of the utility scale by fixing the value of the good to be 1, restricting the value of \( \sigma \) is a substantive restriction. In setting \( \sigma = 1 \), we impose on the model the assumption that the standard deviation of the payoff shocks is equal (in utility terms) to the difference between having and not have the good. Whether or not this assumption is plausible — it most likely is not — it is not one that can be tested in the data. For any \( \sigma > 0 \) that we choose, we can find a set of payoffs that predict the observed distribution of outcomes.

Having fixed \( \sigma \), the mapping \( T_p \) can be restated as

\[
T_p : (\bar{W}_A, \bar{W}_B, \bar{a}) \rightarrow (\Pr(SF), \Pr(BD), \Pr(CD)).
\]

This mapping involves three equations and three unknowns. It is now at least possible that each payoff profile implies a different outcome distribution and, thus, each outcome distribution is supported by a single profile of average payoffs. In the appendix, we show that that \( T_p \) is one-to-one
in this way, so that a mapping, $T_p^{-1}$, exists and its domain includes all possible outcome distributions. Thus, we can estimate the parameters of the model, and those parameters will always predict exactly the observed data.

**Agent error/Quantal response QRE**

Because the QRE equilibrium concept involves inherently stochastic play, the probabilities of each outcome are established directly by the statement of the equilibrium given in Section 2.

As above, we can define a transformation, $T_q$, that maps parameter values into outcome probabilities:

$$T_q : (W_A, W_B, \bar{a}, \sigma) \longrightarrow (\Pr(SF), \Pr(BD), \Pr(CD)).$$

Again we encounter the problem that this transformation defines a system of three equations and four unknowns, and no unique inverse for the transformation exists. Every distribution of outcomes is supported by a continuum of average payoff profiles and shock variances. As in the NE case, we must make additional restrictions to identify the model. In order gain identification of the payoffs, we again normalize $\sigma = 1$. Note further that because the QRE equilibrium definition depends on the fixed $V_A$ and $V_B$, the multiplication of $\sigma$ by a constant cannot be cancelled out by multiplying the remaining payoffs by the same constant. Thus, we expect that fixing $\sigma$ at different levels will effect not only our inferences about the relative values of the estimated and fixed payoffs, but also our inferences about the relative values of the estimated payoffs themselves.

We can then redefine $T_q$, to be

$$T_q : (W_A, W_B, \bar{a}) \longrightarrow (\Pr(SF), \Pr(BD), \Pr(CD)).$$

In the appendix we show that $T_q$ is one-to-one and $T_q^{-1}$ thus exists.

**Incomplete information/PBE**

The description of the PBE mapping from average payoffs to outcome probabilities is complicated by two features. In section 2, the probabilities for each outcome were conditional on the value of the audience cost $a$. As is shown in the appendix, it is straightforward to write down the probabilities of each outcome given the average war costs and the realized audience cost under the assumption that $W_A$ and $W_B$ are distributed normally. The analyst, however, does not observe the
realized value of $a$. Therefore, to find the marginal distribution of the outcomes incorporating the analyst’s uncertainty about $a$, we write

$$\Pr(P|\bar{W}_A, \bar{W}_B, \bar{a}, \sigma) = \int \Pr(P|\bar{W}_A, \bar{W}_B, a, \sigma)\phi(a - \bar{a})da$$

where $\phi$ is the standard normal density function. This integral has no closed form solution and must be approximated numerically. We accomplish this approximation by Monte Carlo integration.

Given this marginal distribution, we can again construct a mapping from $(\bar{W}_A, \bar{W}_B, \bar{a}, \sigma)$ to $(\Pr(SF), \Pr(BD), \Pr(CD))$. As above this mapping involves more unknowns than equations, so we again constrain $\sigma = 1$ in order to identify the payoffs. We then have a mapping:

$$T_b : (\bar{W}_A, \bar{W}_B, \bar{a}) \rightarrow (\Pr(SF), \Pr(BD), \Pr(CD)).$$

We conjecture [at the moment] that this mapping has a global inverse and that the domain of this inverse exhausts the possible probability distributions of the outcomes.

We so far we have established that even with this very simple data structure, we cannot identify all of the free parameters of the model and must make a non-innocuous restriction (setting $\sigma = 1$). We have also shown that any distribution of node outcomes can be supported by each of these models. Thus, without further restrictions, we cannot discriminate between the models on the basis of the observed data. Because the likelihood of the data (1) is maximized when observed outcome frequencies equal the predicted probabilities, the inverse mappings of $T_p$, $T_q$, and $T_b$ are all ML estimators.

4 Results and Implications

In this section, we explore the main implications of these observations. What are the costs of not being able to discriminate between the three models using the observed data? How do the identifying restrictions we were forced to make affect our inferences about the payoffs and, by extension, the effects of covariates on them? The answers to these questions are explored through a series of computational exercises. We first consider the mappings $T_p$, $T_q$, and $T_b$ and their inverses in more detail and show that the three models have very different properties. In several cases,
changing a particular average payoff shifts the frequency of a given outcome in opposite directions depending on which model is assumed. Similarly, observing an increase in the frequency of a particular outcome lead to opposite conclusions about the whether a particular average payoff has gone up or down. We then consider how the need to restrict the values of the good and the variance of the disturbance term influence what we can say about the payoffs and their determinants.

4.1 The Effects of Information Structure

Figure 2 shows how the average payoffs, $\bar{W}_A$, $\bar{W}_B$, and $\bar{a}$, map into outcome probabilities by exploring the comparative-static relationships predicted by each of the three information structures. We assume a baseline payoff vector of $(\bar{W}_A, \bar{W}_B, \bar{a}) = (-2, -2, -1)$ and vary each of the payoffs from -3 to 1 while holding the others constant. The comparative statics for the NE and QRE models can be directly calculated from the node probabilities shown above. In the case of the PBE, we can directly calculate the node probabilities conditional on a given realization of $a$, but we then have to use numerical methods to integrate over $a$. Figure 3 shows similar comparative statics, in this case making the outcome probabilities conditional on a challenge having been made.

The panels show the complexity of some of the functional relationships, as well the similarities and differences in the comparative-static predictions associated with the three information structures. Several of the relationships are quite straightforward and robust to different assumptions. As war becomes more attractive to $A$, in expectation, $B$ is more likely to concede if challenged, and $A$ is more likely to make a challenge. As war becomes more attractive to $B$, in expectation, $B$ is less likely to concede if challenged, and $A$ is more likely to accept the status quo. As it becomes more costly for $A$ to back down from its threat (i.e., $a$ goes down), the probability of that outcome decreases, and the probability that $B$ resists given that it has been challenged also decreases. Even in these cases, though, the exact functional forms of the predicted relationships can vary considerably.

In other cases, the comparative-static predictions associated with the different information structures go in different directions for at least part of the range of parameter values. For example, in the NE, an increase in the expected value for war for either state strictly increases the probability of the $SF$ outcome. This is because war takes place under complete information if and
Effects of changing various game payoffs on
the fraction of times each terminal node is reached

Figure 2: Each panel shows how the fraction of outcomes of the crisis bargaining game that end at particular node vary as a function of the given payoff. Each payoff is varied from -3 to 1. The baseline payoff vector \((W_A, W_B, A) = (-2, -2, -1)\).
Effects of changing various game payoffs on the fraction of times each terminal node is reached conditional on challenge

<table>
<thead>
<tr>
<th></th>
<th>Stand</th>
<th>Back Down</th>
<th>Concede</th>
</tr>
</thead>
<tbody>
<tr>
<td>War cost for A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$W_A$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>War cost for B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$W_B$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audience cost for A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tilde{a}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

QRE = - - - - - - , PBE = - - - - - - , NE = - - - - - -

Figure 3: Each panel shows how the fraction of outcomes of the crisis bargaining game that end at particular node vary as a function of the given payoff. Each payoff is varied from -3 to 1. The baseline payoff vector $(W_A, W_B, \tilde{a}) = (-2, -2, -1)$. 
Effect of changing relative frequencies of each game node on the implied game payoffs

<table>
<thead>
<tr>
<th>Game Node</th>
<th>War Cost for A</th>
<th>War Cost for B</th>
<th>Audience Cost for A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stand Firm</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>Back Down</td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>Concede</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
</tr>
<tr>
<td>Status Quo</td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
</tbody>
</table>

Figure 4: Each panel shows an implied game payoff as a function of the fraction of outcomes that end at a particular node. The baseline distribution of outcomes \( \{P(SF), P(BD), P(CD), P(SQ)\} = \{0.10, 0.25, 0.15, 0.50\} \). The probability of each outcome was varied from 0.1 to 0.9 holding the relative probabilities of the remaining outcomes fixed. For example, when the fraction of cases that end at CD is varied, the ratio of SQ to SF is held at 5:1.
only if \( W_A \) is greater than \( \max(0, a) \) and \( W_B \) is greater than zero. On the other hand, increasing the expected values of these terms can decrease the likelihood of war in the incomplete information game. The reason is that, as the expected values for war become very high, states are less likely to make the "mistakes" that lead to war under incomplete information. Beyond a certain point, increases in the expected value of \( W_B \) make \( A \) less likely to bluff, and increases in the expected value of \( W_A \) make \( B \) less likely to resist. Thus, while the probability of war is always higher under incomplete information, increasing the expected values for war can decrease the frequency of "unwanted" conflicts—i.e., those that would not happen under complete information. Indeed, as these terms get arbitrarily large, the probabilities of war under complete and incomplete information converge.

From the standpoint of estimating the payoffs, of course, what we are really interested in is the inverse mappings — that is, what a given outcome distribution implies about the payoffs. Figure 4 compares the estimates of \( \hat{W}_A \), \( \hat{W}_B \), and \( \hat{a} \) for different outcome distributions. To make this figure, we assume the following baseline distribution: \( P(SF)=0.10 \), \( P(BD)=0.25 \), \( P(CD)=0.15 \), and \( P(SQ)=0.50 \). In each row, we vary the frequency of the specified outcome from 0.1 to 0.9, while holding the relative probabilities of the other three outcomes constant. Thus, for example, when we vary the fraction of cases that end at \( CD \), the ratio of \( SF:BD:SQ \) remains fixed at 2:5:10. The graphs show how the QRE, PBE, and NE estimates of each parameter change in response.

As before, there are some cases in which a change in the distribution of outcomes causes similar reactions by the three different estimators. For example, if the \( SF \) outcome becomes more relatively more frequent, all of the estimators conclude that war has become more attractive to both states. At the same time, there are some very dramatic, and instructive, differences in some cases, particularly in the way the PBE and QRE estimators react to changes in the frequency of \( BD \) and \( CD \) outcomes. As the proportion of back down outcomes increases, all estimators conclude that the audience costs must have decreased (i.e., \( \hat{a} \) increased), but the PBE and QRE estimators reach very different conclusions about how \( W_B \) must have changed. In the PBE, as the expected value of \( W_B \) goes down, more types of \( A \) are tempted to bluff in order to exploit \( B \)'s desire to avoid war. Thus, if \( A \) is caught bluffing more frequently, the PBE estimator concludes that \( W_B \) decreased. With the QRE estimator, on the other hand, a higher frequency of \( BD \) outcomes leads to a higher estimate for \( W_B \). This is because, under the behavioral assumptions of the QRE, states do not bluff
strategically; rather, they do so mistakenly—because their assessment of what they will do in the event of resistance can change from the first node to the third with the new draw of the perceptual disturbance. The best way to account for an increase in BD outcomes, then, is to make backing down more attractive, by increasing \( \bar{a} \), and then increasing B’s incentive to resist.

There are also dramatic differences in how the PBE and QRE estimators react to an increase in the rate of concession by B. When the CD becomes more frequent, both estimators assume that war must have become less attractive for B; however, they disagree about what happened to A’s payoffs. The PBE estimator concludes that the costs of backing down must have increased (i.e., \( a \) decreased), thereby making A’s threats more credible. Such an inference follows naturally from the logic of costly signaling (esp., Fearon 1994b; Schultz 1999). The QRE estimator, on the other hand, concludes that \( W_A \) and \( \bar{a} \) increased in tandem. Such changes increase the probability that A will make a challenge in the first place—thereby giving B many more opportunities to concede—without changing the relative probabilities of the SF and BD outcomes. These contrasting reactions underscore the differences between the behavioral and informational assumptions underlying these two equilibrium concepts. In a QRE, A’s initial move has no influence on B’s choice, other than to determine whether or not its decision node is even reached. To increase the frequency of concession, it is sufficient to increase the rate of challenges. In a PBE, by contrast, A’s strategy influences B’s posterior beliefs and hence its willingness to resist. To increase the frequency of concessions, it is necessary to increase the rate of credible challenges. Thus, in response to an increase in CD outcomes, the QRE estimator tries to make challenges more attractive, while the PBE estimator tries to make challenges more costly, and hence more credible.

All of these results suggest that which information structure one assumes has significant consequences for the estimated payoffs one obtains. While each model can fully account for the observed data, the way each does so can be radically different. Thus, the analyst’s assumption on this score is neither innocuous nor testable. Two analysts with the same model and the same data but different beliefs about the structure of information can arrive at very different estimates of the payoffs—and have no way to determine which set of estimates is better.

Moreover, we can anticipate that the differences between the solution concepts will create addi-

---

8The slight decrease in the estimate of \( W_A \) is needed to ensure that, while \( \bar{a} \) is decreasing, the relative probabilities of SF and BD do not change.
tional problems when covariates are introduced. Consider, for example, a very simple experiment with a control group and a treatment group. Say, for example, that the control group is made up of cases in which state $A$ is nondemocratic and the treatment group consists of cases in which state $A$ is democratic. Such an experiment should permit us to estimate how the regime type of the challenger affects the payoffs. Suppose we find that the relative frequency of $CD$ is higher in the treatment group than in the control group, but that the relative frequency of all the other outcomes remains the same. From this, what could we infer about the effect of democracy in $A$ on the payoffs? Clearly, our inference would depend upon the solution concept employed. If we use the QRE estimator, the increase in the relative frequency of $CD$ will be attributed to an increase in both $W_A$ and $\bar{a}$. We would then conclude that democracy in the challenger makes both war and backing down more attractive. On the other hand, if we use the PBE estimator, the increase in $CD$ will be attributed to a decrease in $\bar{a}$. We would then conclude that democracy in the challenger increases the penalty for backing down. Thus, two analysts with the same data and the same model but different assumptions about the information structure can arrive at radically different estimates of the effects of covariates.

4.2 The Effects of the Identifying Restrictions

Not let us assume that the analyst knows (or is comfortable assuming) which information structure is appropriate. What can be learned from the estimates of the payoff means? As noted earlier, regardless of the model employed, obtaining a unique solution requires not only that we normalize the values of the good, but also that we fix the variance of the random shocks, $\sigma^2$. We asserted earlier that these three restrictions could not be made simultaneously without loss of generality, and we now demonstrate these costs.

The basic problem is that we have set the units of the utility scale in two different ways simultaneously, making it very difficult to attach substantive meaning to any of the estimates. In theory, there are two ways to fix the utility scale without loss of generality. First, one can fix the value of two payoffs and then measure everything relative to their difference. Alternatively, one can fix a single payoff and the spread of the scale. This latter approach is routinely employed when estimating probit models, in which the index threshold and the variance of the error term are fixed.
Effect of varying the value of contested good on the payoffs implied by a given distribution of outcomes

Complete Information/NE  Agent error/QRE  Incomplete Information/PBE

\[ W_A = \quad , \quad W_B = \quad , \quad \bar{a} = \ldots \ldots \]

Figure 5: Each panel shows how value of \( W_A, W_B, \) and \( \bar{a} \) implied by the distribution of outcomes \((0.10, 0.25, 0.15, 0.50)\) vary as a function of the assumed value of \( V_A \).

at zero and one, respectively. Here, we have been forced to do impose both sets of restrictions at the same time. We have fixed two payoffs for each state — the value of not having the good at zero and the value of having the good at one — and we have fixed the variance of random shocks. As we will see these restrictions complicate the interpretation of the payoff means we have estimated.

To see this, we performed additional estimations using our baseline distribution of outcome. Figure 5 shows how the parameter estimates vary as a function of \( V_A \), holding \( V_B \) at one. Figure 6 shows how the parameter estimates vary as a function of \( \sigma \), holding \( V_A \) and \( V_B \) at one. For the purposes of the exposition, these graphs group the estimates obtained from each of the three information structures.

In the case of the complete information/NE estimator, the none of the estimates of vary with \( V_A \); the same would have been true had we varied \( V_B \) instead. This is because, as we saw in table 1, the equilibrium outcome in this world does not depend upon the value of the good to either state, as long as we assume that both are nonnegative. This means that we can say nothing about how the values of war and backing down to \( A \) compare to its value for the good, since they remain the same regardless of what we assume \( V_A \) to be. Similarly, the estimates tell us nothing about how \( \bar{W}_B \) compares to \( B \)'s valuation of the good. Changes in \( \sigma \), on the other hand, do change the
Effect of varying the value of $\sigma$ on the payoffs implied by a given distribution of outcomes

\[
\begin{align*}
W_A &= \ldots, \\
W_B &= \ldots, \\
\bar{a} &= \ldots
\end{align*}
\]

Figure 6: Each panel shows how value of $W_A$, $W_B$, and $a$ implied by the distribution of outcomes (0.10, 0.25, 0.15, 0.50) vary as a function of the assumed value of $\sigma$.

estimated values and in a way that keeps their proportion with $\sigma$ constant. This means that the estimates, and the differences between them, are only meaningful relative to the standard deviation of the disturbance term. As a result, we can have some sense of how large the mean payoffs are relative to the noise in the system, but it is hard to attach any additional substantive meaning to the estimates or their differences. Because the ordering of $W_A$ and $\bar{a}$ are preserved, we can say whether war or backing down is more attractive to $A$ on average, but we cannot interpret the magnitude of the difference in concrete terms.

In the case of the QRE estimator, different assumptions about $V_A$ directly influence the estimates of $W_A$ and $\bar{a}$. Both parameters decrease as the value of the good is assumed to increase. This makes intuitive sense: an increase in $V_A$ makes challenges more attractive, so in order to keep the distribution of outcomes the same, both war and backing down must become less attractive. Moreover, these terms must decrease in a way that keeps the difference between them constant, so that the relative probabilities of $SF$ and $BD$ are preserved. Notice that the difference between $W_A$ and $\bar{a}$ is proportional to $\sigma$ but does not vary with $V_A$. Thus, we can only interpret the difference between these payoffs relative to the noise, but not relative to the value of the good. This means that we can learn nothing about $W_A$ and $\bar{a}$ from estimating the full QRE model that we could not learn by estimating a standard probit model on the sample of resisted challengers — i.e., by looking at
the relative probabilities of $A$'s actions conditional on reaching its final node. A probit model on this decision node would give us exactly what we have obtained here: an estimate of the difference between the average payoffs of backing down and war scaled by the standard deviation of the error term.

The situation becomes even more complicated when we move the PBE estimator, in which case the values we assume for $V_A$ and $\sigma$ influence all three parameter estimates. Here, neither the levels nor the differences of $A$'s payoffs move in proportion with $V_A$ and $\sigma$, making it difficult to know on what scale these payoffs are measured. As before, the orderings of $\bar{W}_A$ and $\bar{a}$ are preserved, so that the most we can learn is which payoff is higher on average. Another troubling observation is that an increase in $V_A$ leads to an increase in the estimate of $\bar{W}_B$. The reason for this is easy to understand: as $A$'s value for the good increases, so does its temptation to bluff, making $B$ want to resist at a higher rate; in order to restore the original distribution of outcomes, $B$'s incentive to resist must be dampened by making war less attractive. This effect confounds our ability to interpret the estimate of $\bar{W}_B$. Recall that, by normalizing $V_B$ to one, we hoped to be able to interpret $\bar{W}_B$ as the mean value of war relative to the value $B$ places on the good. It is clear, however, that the estimate of this parameter hinges on how we normalize $V_A$. With one assumption about $V_A$, we might infer that $B$'s expected value for war is equivalent to $-0.75$ times the value it places on the good, and with a different assumption about $V_A$, we might infer that $\bar{W}_B$ is $-1.25$ times $V_B$. Thus, the scale on which we choose to measure $A$'s payoffs has a substantive effect on our interpretation of $B$'s payoffs. Given that the former is entirely arbitrary, this renders the estimate of $\bar{W}_B$ virtually meaningless.

5 Conclusion

We have shown that estimating payoffs directly from the structure of an extensive form game is a project fraught with severe challenges. In particular, estimates are highly sensitive to untestable assumptions about the information available to the actors. Moreover, they require strong identifying restrictions which make it difficult, if not impossible, to interpret what the estimates actually mean.

We established these points using a simple date generating process in which each payoff was a function only of a constant and a disturbance term. Clearly, researchers interested in testing
hypotheses from IR theory will want to estimate more complicated models that include covariates, such as military power, alliances, regime type, etc. A more rigorous treatment of the issues that arise when covariates are added to the model must await the next version of this paper. The results presented here, however, already anticipate some of the problems that will arise. For example, we saw that different information structures can generate different estimates of the effect of covariates on the payoffs. In addition, all of the problems that arose in identifying and interpreting the mean payoffs will similarly bedevil efforts to estimate the intercept terms of a more complete model. Indeed, it is telling that, in the sole existing effort to estimate an empirical model directly from an extensive form crisis game—Signorino (1999)—all intercept terms were simply dropped. As a result, estimates can depend upon arbitrary scaling choices for the independent variables (e.g., whether democracy is measured on a -5 to 5 or 0 to 10 scale).

What are researchers to do? The arguments in this paper suggest that the goal of estimating payoffs directly from game structures, while laudable, may not be practical or useful. Nevertheless, more limited tests, based on a models comparative-static predictions, can still serve to determine whether the relationships predicted by a formal model are indeed borne out empirically. When a model predicts a monotonic relationship between some parameter and the probability of some observable outcome, and when our theories permit us to operationalize those parameters using measurable covariates, then standard econometric techniques can permit us to assess whether the covariates affect the frequency of the outcome in the predicted direction. This technique can be used in one of two ways. One can assume that the model is correct and then test to see whether the covariate has the hypothesized effect on the payoff by seeing whether it has the right effect on the outcome. Alternatively, one could assume that the covariate affects the payoff and then test whether the model gets the relationship between the payoff and the outcome right.

Nevertheless, the analysis in this paper highlights some general limitations that apply even to this technique. First, because a models comparative-static predictions can depend upon the information structure assumed, these tests may still depend upon the analysts untestable assumption about what the actors know. Thus, researchers need to be sensitive to how robust the models predictions are to different assumptions on this score. Second, by isolating the effect of the covariates on one payoff, one has to take care that the results are not driven by the covariates effect on different payoff.
Consider the following example. The crisis bargaining model shows that an increase in audience costs decreases the probability of resistance by the target (conditional on a challenge). Moreover, this comparative static holds true regardless of the information structure assumed (see Figure 3). Fearon (1994b) argues that democratic states can generate higher audience costs than can nondemocratic states. Combining the comparative-static prediction of the model and Fearon’s theory, one gets a clear empirical prediction: threats made by democratic challengers should be less likely to be resisted. Schultz (1999) found that the prediction was in fact borne out by the data. Such a result has to be considered supportive but not conclusive. First, as Schultz (1999) pointed out, other theories about democracy that do not rely on payoff effects (e.g., greater transparency) also predict a lower probability of resistance. Second, as Figure 3 shows, democracy in the challenger could lead to a lower rate of resistance (conditional on challenge) through its effect on other payoffs. If, as Reiter and Stam (1998a, 1998b) suggest, democracies have better war-fighting ability than nondemocracies, then democracy in A could raise \( \overline{W}_A \) and/or lower \( \overline{W}_B \), both of which would also lead to a lower rate of resistance. Tests using other observable outcomes will not help disentangle these arguments, because the effects of changes in these payoffs are either observationally equivalent or ambiguous.

We do not intend to argue that formal models cannot or should not be tested. Rather, we suggest that nature of strategic interaction limits what we can learn from non-experimental data. If all we know is the relative frequency with which the actors arrive at various outcomes, then it is not a trivial matter to reconstruct the set of payoffs that led them there. The outcome of any given interaction is a product of interdependent choices, shaped both by the individual payoffs and by the information and expectations the actors bring to the game. While game theory provides a set of tools for predicting equilibrium behavior from all of these ingredients, it is not clear that the tools exist for going in the other direction: inferring the payoffs and expectations from the way the game was played.

A Equilibria to Incomplete Information Game

The following describes PBE equilibria for the game presented in Figure 1. The assumptions are as follows:
1. Values for war: $W_A = \tilde{W}_A + \epsilon_A$ and $W_B = \tilde{W}_B + \epsilon_B$ where $\epsilon_A, \epsilon_B \sim \text{i.i.d. } N(0, \sigma^2)$

Disturbance terms are private information of each state.

2. Value of the good is $V_A$ for $A$ and $V_B$ for $B$.

3. Audience costs, $a = \bar{a} + u$ where $u \sim N(0, \sigma^2)$. Audience costs are known to $A$ and $B$ and unknown to the analyst.

A.1 Equilibrium if $0 > a > \frac{-V_A \Phi(-W_A/\sigma)}{1 - \Phi(-W_B/\sigma)}$ and $\frac{\Phi((a-W_A)/\sigma)}{1 - \Phi((a-W_A)/\sigma)} > -\tilde{W}_B + \sigma \Phi^{-1} \left( \frac{-a}{(V_A-a)\sigma} \right)$

State $A$ stands firm at final node if $W_A > a$ or

$$\epsilon_A > -\tilde{W}_A. \quad (3)$$

Let $q$ denote the probability that (3) holds given that $A$ has made a challenge. Assume that the equilibrium takes the following form: $A$ always makes a challenge and stands firm when $W_A > a$; $A$ makes the challenge and backs down if $a > W_A > b$, where $b$ is to be determined; and $A$ chooses the status quo when $W_A < b$. Then, by Bayes’ rule,

$$q = \frac{1 - \Phi((a-\tilde{W}_A)/\sigma)}{1 - \Phi((b-\tilde{W}_A)/\sigma)}. \quad (4)$$

Given $q$, state $B$ resists if and only if $qW_B + (1 - q)V_B > 0$. or

$$\epsilon_B > -\tilde{W}_B - \left( \frac{1 - q}{q} \right) V_B \equiv w^*. \quad (5)$$

Thus, state $A$ expects resistance with probability

$$s \equiv \Pr(\epsilon_B > w^*) = 1 - \Phi \left( -\tilde{W}_B/\sigma - \left( \frac{\Phi((a-\tilde{W}_A)/\sigma) - \Phi((b-\tilde{W}_A)/\sigma)}{1 - \Phi((a-\tilde{W}_A)/\sigma)} \right) V_B/\sigma \right). \quad (6)$$

In equilibrium, it must be the case that an $A$ of type $b$ is indifferent between making the challenge (and backing down in the event of resistance) and choosing the status quo, or $sa + (1 - s)V_A = 0$. This implies

$$s = \frac{V_A}{V_A - a}. \quad (7)$$

Setting (6) and (7) equal to one another, we can solve for $b$:

$$\Phi(b - \tilde{W}_A) = \Phi((a - \tilde{W}_B)/\sigma) + \left( 1 - \Phi((a - \tilde{W}_A)/\sigma) \right) \left( 1 - \tilde{W}_B - \sigma \Phi^{-1} \left( \frac{-a}{(V_A-a)\sigma} \right) \right).$$

For $b$ to be less than $a$, it must be the case that the term in brackets is less than zero. This is true when the first condition above is met. It must also be the case that $b$ is no greater than one, which is true when the second condition is met. Notice that the probability of a challenge,
\[ 1 - \Phi((b - \bar{W}_A)/\sigma). \]

\[
\Pr(W_A > b) = (1 - \Phi((a - \bar{W}_A)/\sigma)) \left(1 - \bar{W}_B - \sigma\Phi^{-1}\left(\frac{-a}{(V_A - a)\sigma}\right)\right). 
\] \hfill (8)

\subsection*{A.2 Equilibrium when 0 > \frac{-V_A\Phi(-\bar{W}_A/\sigma)}{1 - \Phi(-\bar{W}_B/\sigma)} > a}

In this case, the probability that \( W_B > 0 \) is sufficiently high that no types of \( A \) want to bluff. Let \( k \) denote a cutpoint in \([a, 0]\) such that \( A \) makes challenge and stands firm if \( W_A > k \) and chooses the status quo otherwise. Given that all types that make the challenge also stand firm, \( B \) only resists if \( W_B > 0 \), or

\[
\epsilon_B > -\bar{W}_A. 
\] \hfill (9)

Thus, the expected probability of resistance, \( s \), is \( 1 - \Phi(-\bar{W}_A/\sigma) \). For a state of type \( k \) to be indifferent between CH and SQ, it must be the case that \( sk + (1 - s)V_A = 0 \), or

\[
k = -V_A\frac{\Phi(-\bar{W}_B/\sigma)}{1 - \Phi(-\bar{W}_B/\sigma)}. 
\]

Thus, \( A \) challenges if

\[
\epsilon_A > -\bar{W}_A - V_A\frac{\Phi(-\bar{W}_A/\sigma)}{1 - \Phi(-\bar{W}_B/\sigma)}. 
\]

\subsection*{A.3 Equilibrium when \( a > 0 \) or \( \frac{\Phi((a-W_B)/\sigma)}{1 - \Phi((a-W_A)/\sigma)} \leq -\bar{W}_B + \Phi\left(\frac{-a}{(V_A-a)\sigma}\right) \)}

Assume that all types of \( A \) want to challenge, though only those for which (1) holds will stand firm. State B’s posterior belief that (1) holds is the same as the prior, or

\[
q = 1 - \Phi((a - \bar{W}_A)/\sigma). 
\]

Thus, \( B \) resists only if

\[
\epsilon_B > -\bar{W}_B - \frac{\Phi((a - \bar{W}_A)/\sigma)}{1 - \Phi((a - \bar{W}_A)/\sigma)}. 
\]

State A clearly wants to make the challenge if \( a > 0 \). If \( a < 0 \) but the second condition holds, then the expected probability of resistance implied by (11) is less than the critical value in (5) which makes potential bluffers (i.e., those for which \( \bar{W}_A \leq a \)) indifferent between CH and SQ. Hence, they too want to challenge.

\section*{B Proof that all three approaches are consistent with any distribution of outcomes}

\textit{Very Rough Sketch of the proof.} Each of the mappings, \( T_c, T_q, \) and \( T_b \), presented in section 3, takes some

\[(\bar{W}_A, \bar{W}_B, \bar{a}) \in \mathbb{R}^3 \]
That is, each mapping is from $\mathbb{R}^3$ to a simplex on $\mathbb{R}^4$ (a three-dimensional object).

The proof involves showing that for every point in the simplex ($\tilde{P}$) there is one and only one point in $\mathbb{R}^3$ ($\tilde{\mu}$) that is its preimage. The mappings must therefore be one-to-one and surjective. Such mappings are said to be bijective.

To show that a continuous and differentiable mapping is globally bijective, it is sufficient to show that points in the boundary of the domain map to points on the boundary of the range and that the determinant and the Jacobian of the mapping is non-zero for all points in the domain (Buck 1978, p. 361). [STILL WORKING ON A GOOD INTUITION FOR THESE CONDITIONS].

The first condition is fairly easy to show. For each method the following limits obtain:

$$\lim_{W_A \to -\infty} \Pr(BD) = 0$$
$$\lim_{W_B \to -\infty} \Pr(CD) = 0$$
$$\lim_{a \to -\infty} \Pr(SF) = 0$$
$$\lim_{W_A \to 1} \Pr(BD) = 0$$
$$\lim_{W_B \to 1} \Pr(CD) = 0$$
$$\lim_{a \to 1} \Pr(SF) = 0$$

holding the other payoffs fixed. Similar limits demonstrate that some probability approaches 0 as any pair of payoffs goes to infinity or as one goes to infinity while the other goes to negative infinity. Finally, some probability goes to zero if all three payoffs approach some limit. Thus, points on the boundary of the domain map to points on the boundary of the range. Since all points on the boundary of the simplex involve some probability going to 0, these limits establish that points on the boundary of domain are mapped to the boundary of the simplex.

The second condition is more difficult to demonstrate. For the QRE and NE cases, it is straightforward to show that the mapping is everywhere differentiable because the component functions are themselves differentiable. In the PBE this is more difficult.

For the QRE and NE cases, the determinant of the Jacobian can be solved directly and shown to be non-zero basically because it involves sums of products normal cumulative and density functions (sums of products of strictly positive terms). This calculation for the PBE is harder. At present, we can only conjecture that the same condition holds in that case as well.
References


