A Universal Test of an Expected Utility Theory of War

Author(s): D. Scott Bennett and Allan C. Stam

Source: *International Studies Quarterly*, Sep., 2000, Vol. 44, No. 3 (Sep., 2000), pp. 451-480

Published by: Wiley on behalf of The International Studies Association

Stable URL: https://www.jstor.org/stable/3014007

International Studies Quarterly (2000) **44**, 451–480.

A Universal Test of an Expected Utility Theory of War

D. Scott Bennett

The Pennsylvania State University

AND

Allan C. Stam

Dartmouth College

Bueno de Mesquita and Lalman's version of an expected utility theory of war has become one of the most widely cited theories of international conflict. However, the testing of the theory has lagged behind its theoretical development. In its most sophisticated formulation, the theory has been tested on only 707 dyad-years, all drawn from Europe between 1816 and 1970. We present a test of the expected utility theory of war (as developed in *War and Reason*) on the international system from 1816 to 1984. Specifically, we examine the relationship between the main equilibrium variables derived under the "domestic variant" of the international interaction game and behavioral outcomes using multinomial logit. We find that the equilibria correlate with actual behavior in both the set of all dyads and a subset of politically relevant dyads, even after including a set of control variables. The relationship is somewhat less clear among the population of all interstate-directed dyads, however, with key equilibrium variables having smaller effects at increasing the odds of interstate conflict among all dyads. We also present a new software program, *EUGene*, which generates expected utility data and can serve as an important data management tool for international relations researchers.

Bueno de Mesquita and Lalman's variant of the so-called expected utility theory of war has become one of the most important theories of international conflict. Following its most recent explication, a more appropriate label might be "the *War and Reason* game-theoretic theory of war." In *War and Reason* Bueno de Mesquita and Lalman (1992) explicitly model game-theoretic interactions between states, but of course the "International Interaction Game" (hereafter, IIG) in *War and Reason* represents only one of many possible and plausible games of international conflict. Whatever the label, the empirical testing of expected utility theory has lagged behind its theoretical development in many fields, including international relations (Green and Shapiro, 1994). Bueno de Mesquita and Lalman tested their game on

Authors' note: The order of the authors' names is alphabetical and is not intended to connote principal authorship. This material is based upon work supported by the National Science Foundation under Grant No. SBR-9601151. We would like to thank Bruce Bueno de Mesquita for his invaluable assistance in generating expected utility data. *EUGene* is available as freeware, and may be downloaded from our website at http://www.eugenesoftwave.org. The full dataset analyzed here may be generated by that program, or is available from the authors.

Published by Blackwell Publishers, 350 Main Street, Malden, MA 02148, USA, and 108 Cowley Road, Oxford OX4 1JF, UK.

only 707 dyad years drawn from Europe between 1816 and 1970. Testing has been limited because the necessary data for wider analysis—namely, risk attitude scores and utility values for all states and years—have not been available. In the absence of a broad test, the robustness and generality of the theory remains open to question. In an earlier paper, we presented a cross-validation of the tests in *War and Reason* using the population of interstate dyads from 1816 to 1984, following Bueno de Mesquita and Lalman's methods (Bennett and Stam, 2000a). In this paper, we present a more thorough test of the IIG's predicted conflict outcomes, using a more complete dataset that allows us to employ a multinomial logit model to examine a number of hypotheses simultaneously and include a set of controls for other common explanations of conflict discussed in the literature. We find mixed support for the theory across the population of all dyads and politically relevant dyads, and across analyses with and without controls. We also find somewhat different fit between individual equilibrium predictions and conflict outcomes than Bueno de Mesquita and Lalman reported in *War and Reason*, with the weakest support coming for the prediction of negotiation.

Expected Utility Theory

Rational choice applications to war initiation begin with the assumption that states can be modeled as rational actors who make choices about war and peace by assessing the costs and benefits of alternative actions. When faced with a choice between alternatives, leaders will take the option that promises, on average, the best return, or in other words, the highest expected utility. Expected utility theorists have made very strong claims about the theoretical superiority of the rational choice approach based on its cumulative and rigorous applications, and have also made claims about its suitability for policy analysis (Bueno de Mesquita et al., 1985). Unfortunately, in most cases empirical testing has lagged behind the theoretical development of this class of models (Green and Shapiro, 1994).

The application of rational choice to international conflict in its broadest empirical sense is probably most closely associated with Bruce Bueno de Mesquita and David Lalman. Bueno de Mesquita's *The War Trap* (1981) served as a milestone in the application of rational choice theory to the quantitative study of conflict. *The War Trap* remains widely cited in the literature, and has spawned a virtual cottage industry of rational choice scholarship and comment thereon. Bueno de Mesquita's measures of interests, risk, and utility have become the *de facto* academic standard for these concepts in international relations. Initially, the theory was decision-theoretic in nature, with the relevant utility values for empirical analysis limited to two quantities, the utility of war and the utility of peace. In *War and Reason* (1992), Bueno de Mesquita and Lalman expanded the decision-theoretic structure to a game-theoretic one, taking into account the possible strategic interactions of potential belligerents.[1] The IIG posits a series of interactive decision paths that lead to a set of eight different possible outcomes for any dyadic relationship at a given moment. Given adequate data, a prediction can be made of the expected outcome in equilibrium from any given interstate dyadic interaction. The prediction may then be compared to historical situations to ascertain whether the predictions match actual behavior.

Bueno de Mesquita (1981:134–140) argues explicitly that the rational choice models he presents should hold across space and time:

> To be sure, foreign policies in some regions—notably the middle east—have yielded considerably more opportunities for war or serious conflicts than have

[1] Important improvements in the measurement of risk attitudes and utilities were also made between the publication of *The War Trap* and *War and Reason.*

> foreign policies in other regions (by containing a higher proportion of dyads with positive expected utility). Still, the propensity to require the same conditions has not varied meaningfully from region to region or from time to time.

Bueno de Mesquita maintains that while the foreign policies of states may vary, the theory should still fit equally well around the world and across various historical eras. Although the incidence of war tends to vary by region and across time, this variation simply reflects different utility levels rather than differences in foreign policy tendencies or decision-making styles that somehow might be inherent to states in a particular region or time. In the Middle East, as Bueno de Mesquita notes, states may have quite aggressive foreign policies, which in turn lead to more situations in which his so-called war trap condition holds. However, once a crisis is under way, Bueno de Mesquita claims that Middle Eastern dyads will show the same relationship between the utility of war and its outbreak as dyads elsewhere. While components of the utility model (e.g., risk attitude or the utility of war) may vary systematically across regions, this does not imply a different fit to the theory. In *The War Trap*, an analysis of the relationship between expected utility and war split by region and time period supports this argument. In addition, Bueno de Mesquita presents strong empirical support for the broader theoretical arguments developed there, with measures of utility correlating highly with both dispute escalation and the outbreak of war.

While Bueno de Mesquita and Lalman present strong empirical support for the arguments developed from the IIG in *War and Reason*, the scope of the tests differs considerably from that in *The War Trap*. The tests in *War and Reason* were limited to 707 nation-year dyads out of a population of hundreds of thousands of dyad-years around the world between 1816 and 1970. While the sample size is large enough to minimize concerns about sampling error, these 707 cases were all drawn from Europe. The primary reason for the limited analysis was that updated risk attitude scores, an essential component for estimating nations' utility for war, were available only for this subset of countries due to the computational demands of creating these data (Bueno de Mesquita, 1985). Unfortunately, because their analysis was restricted to Europe, earlier arguments about regional and temporal invariance could not be tested with the most recent and more sophisticated version of the theory and data. We believe that for this theory to be more thoroughly tested, it must be tested in the full domain for which it is appropriate; namely, in all regions and times. Given concerns about whether the IIG explains new facts above and beyond those predicted by other theories, it is also important to test the predictions in a multivariate model that includes control variables.

The lack of broad tests of expected utility theory is also important in the context of recent more general criticisms of rational choice theory. Criticisms of rational choice theory in international relations have been primarily theoretical, often questioning whether decision-makers can be considered rational (Lebow, 1981; Jervis et al., 1985) or whether expected utility is as well specified and encompassing as its proponents claim (Simowitz and Price, 1990). Some critics have been willing to make quite strong condemnations of rational choice models. For instance, Green and Shapiro (1994:6) argue, "the case has yet to be made that these models have advanced our understanding of how politics works in the real world." However, given the vigorous theoretical debate, we find it surprising that there has been limited *empirical* criticism of the expected utility theory of war.[2] While scholars find instances of behavior where leaders appear to be

[2] We do not find fault with the argument that better theory is needed in the field of international relations and in particular within the rational choice paradigm. Rather, we concur with this argument, but also feel that serious efforts need to be paid to the empirical testing of previously developed theory as well. This is the task we lay out in this paper.

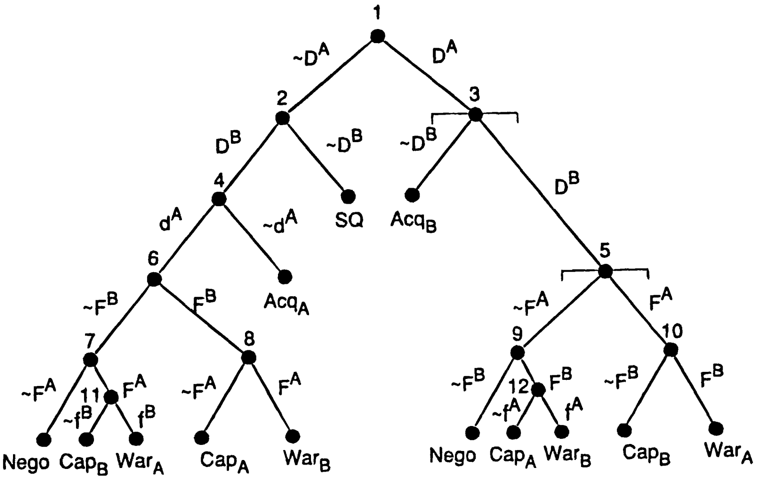

FIG. 1. International interaction game (Bueno de Mesquita and Lalman, 1992).

motivated by "nonrational" concerns or changed their minds over time in ways that are supposedly "nonrational" (e.g., Jervis, Lebow, and Stein, 1985; Kaufmann, 1994), there has been little effort by critics to replicate or extend Bueno de Mesquita's (1981) or Bueno de Mesquita and Lalman's (1992) empirical testing.[3] By offering a more encompassing set of tests and developing the data crucial to conducting extended empirical analyses of the expected utility theory of war, we hope to contribute to the ongoing debate about the theory's usefulness in international relations.

The International Interaction Game: Equilibria as Predictors

The critical variables for testing the expected utility theory of war as applied most recently in Bueno de Mesquita and Lalman (1992) are the game-theoretic equilibrium predictions that emerge from the IIG under its complete-information domestic politics variant.[4] Making a prediction of the equilibrium behavior in a given case in turn depends on a variety of input data on national capabilities and alliances, which combine to create estimates of states' utility for outcomes, which in turn lead to equilibrium predictions.

Figure 1 presents Bueno de Mesquita and Lalman's international interaction game. Bueno de Mesquita and Lalman provide a complete discussion of the structure of the IIG and the choices in it, which we will not duplicate here. A very general overview is that the game represents an interaction structure between a potential conflict initiator (state A) and a potential target (state B). The game begins with a choice by state A to issue some type of demand to state B. If A does not make a demand, then B may make a demand; if neither does, then the outcome of the interaction is the status quo. If one state does initiate a crisis by making a demand, then at subsequent decision nodes in the game actors A and B alternately make choices about whether to make further demands (escalating

[3] Signorino (1999) and Smith (1999) present different estimation procedures but they use the same data as Bueno de Mesquita and Lalman.

[4] Bueno de Mesquita and Lalman lay out several different variants of the IIG. Each variant is dependent on different assumptions about preferences and payoffs across the eight possible outcomes of the game. The most general analysis takes place on the domestic politics variant.

the conflict) or not (resulting in some type of settlement). Each actor makes choices that appear to yield it the greatest expected utility when it looks ahead to the outcome. The key for purposes of discussing the empirical application of the game is to note the eight possible game outcomes (eight terminal nodes). Each of these nodes represents a potential final situation in which the members of a dyad might find themselves after they complete their interaction, and each has some analogue in the real world that can be measured. The eight possible outcomes to the game are (1) a status quo outcome (no dispute, designated SQ in the figure); (2) a challenge resolved by negotiation (Nego); (3) a challenge resolved by state A's acquiescence (giving in, designated $\mathrm{Acq_A}$); (4) a challenge resolved by A's capitulation (an acquiescence to B's demand but under the threat of force, $\mathrm{Cap_A}$); (5) a challenge resolved by B's acquiescence ($\mathrm{Acq_B}$); (6) a challenge resolved by B's capitulation ($\mathrm{Cap_B}$); (7) a war[5] initiated by A ($\mathrm{War_A}$); or (8) a war initiated by B ($\mathrm{War_B}$). Each state expects to receive some utility at each terminal node, resulting in a game with 16 relevant utility values.

The states' utilities for the 8 possible outcomes combine in a nonlinear manner to yield a prediction of what the game's outcome should be in equilibrium. Under conditions of complete and perfect information, the game can be easily solved using backward induction. This is done by working backwards up the game tree, determining at every choice node which of two options provides a higher payoff, given the actors' preferences and the potential choices that were anticipated further down the tree. Bueno de Mesquita and Lalman further focus on specifying a series of assumptions to divide the game into a "realpolitik variant" and a "domestic politics" variant. Under the realpolitik variant, war is never expected to occur under complete information (this is consistent with Fearon's 1995 argument that war is never rational under complete information conditions), and is not analyzed further. Fuller attention is given to the domestic politics variant, where a variety of outcomes is expected given different input conditions. Specifically, Bueno de Mesquita and Lalman concentrate on developing a series of propositions that specify what combinations of utility values for the two players yield each outcome in equilibrium (1992:72–92). Given empirical estimates of states' utilities in a given year, we can determine this expected equilibrium outcome and compare it to actual events to assess how well the prediction correlates with actual behavior. Examining these equilibrium predictions yields the broadest possible test of the expectations of the model for general behavior.[6]

While an interaction between states in the IIG is dyadic, such an interaction is also *directed.* In any dyad A vs. B, the identity of the potential initiator under consideration has an effect on the expected outcome. Given that the utility for outcomes is likely to be different between state A and state B, a different equilibrium outcome could be predicted for the dyad A vs. B and the dyad B vs. A in any given year. Consequently, game-theoretic equilibrium predictions must be made for a *directed* dyad-year, our unit of analysis in the empirical tests. Our dependent variable will take the form of directed-dyad dispute outcomes.

In general, we expect in our analysis that the prediction of the status quo equilibrium will have a positive correlation with the occurrence of the status quo,

[5] Technically, the game predicts up to the mutual use of force. Bueno de Mesquita and Lalman label this outcome "war," but this does not imply that other conditions for war (such as sustained fighting and a high number of casualties) are met. Theoretically, there is no separation between the mutual use of force and a larger war in the game.

[6] Modelers often use the complete information version of a game only as a stepping-stone to versions with incomplete information. Given the difficulty of measuring beliefs in an incomplete information setting given a dyad-year analysis, we have not done this, and focus instead on whether the equilibrium predictions under complete information correlate with actual behavior. Given additional measures, it would be possible to expand tests to include other hypotheses under incomplete information.

but a negative relationship with other levels of escalation. We would expect the acquiescence and negotiation equilibria to correlate positively with dispute initiation and escalation to medium levels, since they represent middle levels of escalatory behavior, but negatively with escalation at very high levels. Finally, we would expect that the war equilibrium should predict both a challenge (the initiation of a militarized dispute) and the escalation of the dispute through all levels of force, up to and including the most serious levels of violence.

The rest of the paper proceeds in three main sections. In the next section, we provide the details of our research design, discuss our population and samples of cases, define the dependent variable, and in particular explain the development and measurement of the independent variables used to test the theory. In the second remaining section, we present our multivariate analyses examining the relationship between the IIG equilibria and actual state behavior. Finally, we conclude by summarizing the areas where we find support for the arguments of *War and Reason* and areas where we do not, and discuss our plans for future research in this area.

Research Design

a. Population of Cases

The directed dyad-year is our unit of analysis throughout this project. A directed dyad is a dyad in which direction is specific, separating (for instance) the Germany vs. Russia dyad from Russia vs. Germany. Because the IIG may predict different equilibria for the directed-dyad A vs. B compared to B vs. A, and because the independent equilibrium predictions depend on data available annually (capability and alliance data), the directed dyad-year is the appropriate unit. The IIG predicts all of its equilibria from a baseline situation of no conflict, that is, from a status quo situation. We thus must start with the set of cases that could have status quo outcomes, namely, the population of dyad-years, rather than starting with a subset of cases already in dispute.

We conduct analysis of both the set of "politically relevant" directed dyad-years, and the population of all directed dyad-years from 1816–1984.[7] A critical issue to consider when applying any model of international decision behavior to actual cases is where and when state leaders actually match the specifications of the model, if ever. In the case of testing predictions from the IIG, the critical assumption is that the states under consideration are indeed interacting according to the choices assumed by the game. It can be argued that both politically relevant dyads and all dyads constitute samples where the IIG and related tests should apply. Politically relevant dyads are those dyads in which the two states are contiguous, plus all dyads where at least one member is a major power. Maoz and Russett (1993) argue that these dyads constitute a reasonable set of baseline dyads in which conflict is a possibility. They justify this assumption on the observation that most states do not have the power projection capabilities to initiate conflict against states to which they do not have direct land access. These are dyads where states are likely to have interests vis-à-vis one another, and so where they are particularly likely to interact following the rational pattern predicted by the IIG. However, we also examine the subset of all dyads. While the terminology of "politically relevant" dyads suggests that most other dyads are simply irrelevant (insofar as predicting conflict) and can therefore be excluded from analysis, 20–25 percent of all militarized disputes actually occur among

[7] 1984 is the limiting final year in our analysis because Correlates of War alliance data (necessary to develop the measures of expected utility that go into our key independent variables) only runs through 1984. When the revised version of the alliances data is available, they will be included in the data produced by *EUGene*.

"irrelevant" dyads (depending upon how contiguity is defined). If we focused only on the subset of politically relevant dyads, we would have no opportunity to explain those disputes. Moreover, rational choice scholars make no caveats that their data or theories should apply only to some subset of cases defined *ex post* as relevant. As specified, we should expect the theory to hold in any dyad, even in dyads without frequent interaction.[8] The expected utility measures we use capture the drop-off in military capabilities over long distances and so reflect the ability or inability of certain states to gain utility from undertaking conflictual activities. Similarly, if states are far apart and have neither common nor conflicting interests, then the utility to be gained or lost in an interaction is near 0. If either of these cases holds, then we would expect to find support for the expected utility theory of war even among dyads that have little interaction, because we should predict status quo outcomes in many international relationships. At a minimum, if our findings do not support the universal application of the IIG and associated measures from Bueno de Mesquita and Lalman (1992), they represent an important caveat and critique (as anticipated but not tested by constructivist and psychological critiques of rational choice models) when applying the theory to international behavior.

After starting with the set of politically relevant or all dyad-years, we drop dyad-years in which a militarized dispute was ongoing at the beginning of the year because leaders typically do not have an opportunity to engage in a new strategic interaction if they are already involved in a dispute.[9] Thus, we do not code ongoing disputes as new militarized disputes as do Oneal and Russett (1997). The IIG makes no predictions about the continuation of conflict, but only about their initiation and initial escalation. In addition, we believe more generally that it is inappropriate to code decisions to continue a conflict as equivalent to decisions to start it or to escalate to some level. We also drop the directed dyad B–A (target vs. initiator) for dispute dyads where A initiates against B, since when A initiates against B, B typically does not have another opportunity to initiate against A.[10] We include both the directed dyad A vs. B and B vs. A for dyads with status quo outcomes, since we are certain that both directions had no dispute initiation.

We also drop dyads involving "joiners" in our analyses. "Joiners" are states that become involved in a MID but not on day 1, and so are distinct from conflict initiators who start a MID by becoming involved at the outset. We code as "true" initiators only the states on the initiating side (side A) on day 1 of the MID, and as "true" targets only the states on the target side of the MID involved on day 1. We give a non–status quo dependent variable value to the set of all combinations

[8] This is simply an extension of the assumption that the expected utility conditions should fit equally well in all regions and periods. Some scholars might argue this is not possible. Constructivists such as Wendt argue, for example, that state interactions can give rise to new and mutually evolving identities for particular dyads, which in turn may serve to either ameliorate or exacerbate the so-called war trap conditions.

[9] The exception is that we include dyad-years where a dispute was ongoing if there is also a new dispute initiation in that year, which indicates that the first dispute ended in time for leaders to make another decision regarding conflict. For an extended discussion of the effects of these and other sampling decisions, see Bennett and Stam, 2000b.

[10] Again, the exception is that we include directed dyad-years B vs. A even when A initiated against B if B initiates a new dispute in that year, which indicates that the first dispute ended in time for leaders to make another decision regarding conflict. We normally drop B vs. A (although the international interaction game allows us to study the outcome of a crisis from either A's or B's perspective) because of censoring. In a year in which A has initiated a dispute against B but B did not also initiate a dispute against A, the outcome of dyad B vs. A is actually censored. We do not know whether, if A had not initiated against B, B *would have* initiated against A. The noninitiation of B vs. A does not necessarily imply a status quo outcome, but neither do we observe an initiation and so one of the other seven outcomes to the B vs. A directed dyad. Since it is censored, we cannot normally include an observation for B vs. A in years in which A initiated against B (there are 75 cases where A and B initiate against each other in the same year where both directions are included).

of "true" initiators against "true" targets. We drop joiners for two reasons. First, we believe that cases where behavior is related to a true, new dispute initiation and escalation best reflect the IIG and empirical measurements. In many cases, joiners enter a dispute well after it has begun (even years later), and often after the outcome (highest hostility level) has been determined. For example, when Rumania and Greece chose to enter WWI in 1916 and 1917, respectively, the highest hostility level of the conflict (full-scale war) had already been reached, and so we do not consider their actions to contribute to a test of the IIG. Simply, their strategic game was not the same as that considered by Serbia, Austria, and the other initial participants in WWI. In addition, later decisions to join an ongoing conflict rely on different information and expectations about the future than the states initially considering beginning the conflict. Simply put, the level of uncertainty faced by joiners is fundamentally different from that faced by the conflict initiator(s). While excluding joiners undercounts subsequent interactions in large events such as the world wars, those interactions require a different game tree for theoretical analysis and a different measurement for empirical testing.[11]

Our second reason to drop joiners is that in fact there is not enough information in the COW MID dataset to construct truly directed dyads involving joiners. The MID data tell us which side of a dispute is the initiating side; we assume that all states involved on day 1 actively chose to initiate against the target side. However, for states that become involved in subsequent years, while we do know what side they were on (initiator's or target's side), we do not know if they *chose* to join the dispute or if they were in fact targeted by one of the states already involved. "Joiner" is really a shorthand term for "states that became involved later" and carries problematic connotations. So for instance, the MID data can only suggest that in the Korean War the United States is a "joiner" on the "target" side, but does not indicate whether the U.S. actively chose to intervene or became involved only because it was attacked. Because we cannot know this from the data, we would drop the U.S.–North Korean directed dyads.

Between 1816 and 1984, we have adequate data on capabilities and alliances to generate expected utility information for 810,900 out of 812,426 directed dyad-years. After dropping cases where disputes were ongoing at the start of the year, and after dropping the target vs. initiator directed dyads, there are 806,391 cases. Dropping joiners to the target side leaves us with 805,910 observations for our initial analysis. Of these, 2,278 are directed dyad-years with non–status quo outcomes. This skewed distribution makes some analyses difficult, as we are predicting quite rare events from it. However, it is by generating and using this full set of cases that we can best assess whether the predictions of the IIG correspond to actual dispute and war outcomes around the world and over time.

b. Dependent Variable: Occurrence and Escalation of Militarized Interstate Disputes

To analyze the fit of the IIG to actual events, we code a dependent variable that marks the occurrence of conflict outcomes paralleling those of the stylized game. We use data on militarized interstate disputes to indicate which of the game outcomes best captures the actual interaction in a particular dyad. The Correlates of War MID dataset contains data on all hostile interactions between states from 1816 to 1992 that include a threat, display, or use of military force by at least one participant. Using v2.1 of this dataset (Jones et al., 1996), we focus on the highest level of hostility reached by each state in a dispute. Within the MID

[11] Note that these "joiners" were often included in Bueno de Mesquita and Lalman's analysis, as they included all dyads in a dispute where at least *one* state was part of the dispute on the first day. This is one source of difference between our results and theirs. As we argue above, joiners and initiators need to be analyzed separately.

data, a level "1" hostility score indicates that a state took no militarized action, a "2" indicates that a state threatened to use military force, a "3" indicates a show of force, a "4" indicates the use of force, and a level "5" indicates a war. We take the presence of a militarized dispute in the first place (which requires that at least one side reached at least a level 2) as indicating a challenge to the status quo.[12] After such a challenge is made, we examine the combination of final hostility levels to determine what game outcome best describes the resulting escalation. Following Bueno de Mesquita and Lalman (1992: Appendix 1), we coded cases where the highest hostility levels showed that both states either used force (level 4) or went to war (level 5) as fitting into the "War" outcome of the game (more appropriately referred to, we believe, as "mutual force").[13] In an actual interaction between two states A and B, cases were coded as "Negotiation" when both sides escalated to an equal hostility level above level 1 but both remained below a level 4. "Acquiescence by A" was coded when B's level of hostility was greater than A's, but B's was still less than 4. "Capitulation by A" was coded when B's level was greater than A's, and B's was level 4 or higher. Similarly, "Acquiescence by B" was coded when A's level was greater than B's but A's was less than 4, while "Capitulation by B" was coded when A's level was greater than B's and A's was level 4 or higher. Cases where no MID was initiated by one state against the other–and hence where no initial challenge is ever made–were coded as falling into the "Status Quo" outcome. Because there are only 9 cases after dropping joiners in which an acquiescence by A was the outcome (too few to analyze separately), we then recoded acquiescence by A outcomes into the status quo category.

After coding the interaction in each directed dyad-year as corresponding to one of the 8 outcomes, we combined the individual outcomes into a single categorical variable representing the level of escalation reached in the dyad. The dependent variable then combines the presence of both an initial demand and dispute escalation down the game tree. Our dependent variable has 6 categories. Category "0" represented the status quo, category "1" an acquiescence by B, category "2" a negotiation, "3" a capitulation by A, "4" a capitulation by B, and "5" the mutual use of force or war. Because these outcomes are unordered in terms of the game structure and independent variables (utilities do not relate monotonically to any outcome category, and the game does not pass through a consistent or logical progression of these outcomes on its way to a final outcome), we use multinomial logit.

One possible source of differences in our results from those of Bueno de Mesquita and Lalman (1992) lies in our use of the newest version (v2.1) of the COW MID data to identify disputes and dispute outcomes. The new MID dataset contains many more disputes than the original MID data and also has recoded many original disputes. Table 1 presents a cross-tabulation between the dispute escalation levels in the new and old MID dataset for the 1,414 dyads that were analyzed in *War and Reason* (707 dyads, with two participants in each dyad). While the outcome codings are highly correlated ($r = 0.73$), there are important differences. Note in particular the many cases where disputes had been coded (old hostility level > 0) that are now coded as non-MID dyads. These differences in turn lead to differences in the coding of outcome categories (acquiescence,

[12] By using MIDs, we assume that all demands fitting the initial demand in IIG are accompanied by a threat of force and so are captured in the MID data, and are thus testing the predictions of the IIG on a subset of relatively high intensity demands. If we had data on "pre-MIDs," that is, demands and interactions before the threat of force, we could use this data to examine the fit of the model at an even more preliminary level of demand.

[13] Because the IIG only predicts escalation to the level of mutual force, and not additional escalation to war once force is used, we cannot make separate predictions of the reciprocated use of force vs. full-scale war. It is also impossible for the game to predict a war started by the potential target (War_B) under complete information, and so we do not distinguish wars started by the dispute initiator A from those started by the target B.

TABLE 1. Cross-Tabulation of Hostility Levels, Old MIDs in Bueno de Mesquita and Lalman, 1992, vs. New MIDs

		Hostility Level, New MID Data						
		0	*1*	*2*	*3*	*4*	*5*	*Total*
Hostility Level, War and Reason MID data	0	476	0	0	0	0	0	476
	1	35	86	2	12	12	7	154
	2	19	1	26	7	6	2	61
	3	60	15	4	111	13	6	209
	4	57	28	1	23	190	5	304
	5	35	3	2	0	10	160	210
Total		682	133	35	153	231	180	1414

TABLE 2. Cross-Tabulation of Hostility Levels, Initiators vs. Initial Targets, All Dyads, 1816–1984

		Initiator Hostility Level				
		2	*3*	*4*	*5*	*Total*
Target Hostility Level	1	111	322	599	3	1035
	2	3	23	25	0	51
	3	9	121	96	1	227
	4	4	62	509	10	585
	5	0	2	16	83	101
Total		127	530	1245	97	1999

status quo, war, capitulation, and negotiation).[14] Table 2 shows the final relationship between initiator and target levels of hostility for the dispute dyads analyzed here; this table can be compared to Bueno de Mesquita and Lalman, 1992:285.

c. Method of Analysis: Multinomial Logit

We employ multinomial logit with a standard error correction for clustering within dyads to analyze the relationship between the IIG equilibria and the categorized conflict outcomes.[15] Multinomial logit is the appropriate technique to use to analyze categorical variables in which there are multiple unordered outcomes, or where the independent variables are expected to have a non-monotonic effect across categories.[16] It estimates the probability that the actual

[14] In this table, a value of "0" indicates that no militarized dispute occurred, while a value of "1" indicates no militarized response to an initiation by another state. Both imply that the state took no militarized action.

[15] Conventional standard error calculations assume that the observations in the dataset are independent and that the error variance is constant (see Beck, 1996, and Huber, 1967, for discussion of the problem and White, 1978, for the correction).

[16] Signorino (1999) and Smith (1999) suggest other methods to be used when statistically estimating models with strategic interaction. Because we have resolved some anomalies in the creation of IIG equilibria, we do not have to use these methods. In particular, Signorino addresses two issues in comparing the subgame perfect equilibria from the IIG to actual outcomes. The first concerns the way in which the 2x2 tests performed by Bueno de Mesquita and Lalman may conceal a large amount of prediction error in the IIG. Because we keep outcomes disaggregated in our multi-chotomous dependent variable, and because we include each equilibrium prediction as a separate dummy variable in a single estimation, this is not a problem for our analysis. The second concerns unavailable cost data and nonassessable equilibrium conditions from the IIG. Bueno de Mesquita and Lalman acknowledge that they do not have a method for estimating a set of three cost terms α, τ, γ, and so "do not distinguish between them" (1992:298). In the empirical generation of utility scores, this meant that in fact they assume that the unknown parameters α, τ, γ take value 1 (Bueno de Mesquita, personal communication). Given an assumed parameter value of 1, we have

outcome Y will take on each of a set of discrete possible outcomes, given a vector of independent variables X.[17] Given J+1 outcomes, J equations (sets of coefficients β) are estimated which show the effects of variables on producing a particular outcome as compared to the base case (which we will take as the status quo). Estimates are relative to a base category (the selection of the base category is mathematically unimportant). Then, given J+1 outcomes numbered 0, 1, 2...J; we obtain the predicted probability that any given case (set of data x_i) will have a particular outcome Y as follows:

$$\text{Probabiliy (Y = j)} = \frac{e^{x_i\beta_j}}{1 + \sum_{k=1}^{J} e^{x_i\beta_k}} \text{ for j = 1,2...J.}$$

$$\text{Probability (Y = 0)} = \frac{1}{1 + \sum_{k=1}^{J} e^{x_i\beta_k}}.$$

In our case, using multinomial logit will allow us to estimate independently the effects of our independent variables (equilibrium predictions) on whether the dispute ends up at each outcome category. We can assess how the game equilibria correlate with the initiation of disputes (by examining whether they predict any outcome level other than the status quo) and how far they proceed through the IIG tree. This technique will also allow us to obtain a nuanced picture of how our independent variables influence conflict outcomes. Using multinomial logit with our dependent variable and the population of all dyads also helps us to avoid the selection bias that would be present if we began an analysis of escalation up the game tree starting only with the cases in which a dispute had been initiated (e.g., Fearon, 1995). Thus, we are able to get useful leverage in a variety of ways by using cases outside of just those with disputes.

Because a number of coefficients are produced for each substantive independent variable, it is difficult to interpret whether variables matter in the sense of being statistically significant and what their substantive effects are. Even given

adequate information to solve the game and make a unique point prediction (subgame perfect equilibrium) following Bueno de Mesquita and Lalman's methods for each dyad that fully takes into account the preceding strategic interaction. In turn, given this single dummy variable prediction *which already takes into account the strategic interaction in the game*, we can assess whether or not the prediction coincides with a more frequent occurrence of the corresponding outcome without building a strategic choice estimator. [The one exception where we do not have a unique equilibrium outcome involves estimation of the War_B equilibrium; here, there are two ways to code outcome predictions, and the two alternatives made no substantive difference in our empirical estimations. See footnote 24.]

More generally, Signorino's methods are particularly appropriate when we wish to regress the components of a strategic choice problem (e.g., the utility scores or quantities related to probabilities of different outcomes) on outcomes (e.g., dispute or war initiation). Often, rough indicators of these utilities and probabilities are all that analysts have when trying to move toward an empirical prediction of conflict. Signorino gives the example of modeling utilities as being based on some function of military capabilities, other assets, and shared democracy. When these utilities and probabilities combine in an interactive or nonlinear fashion (as they typically do in a strategic interaction), it is often inappropriate to simply include the individual variables in a linear additive fashion. In such a case, we should instead use logit quantal response equilibria (LQRE) to derive strategic choice outcome probabilities derived from the relevant game tree (it is important to note that we need to know the structure of the game to be able to specify these probabilities), or use Bayesian methods such as in Smith, 1999. With properly specified probabilities, we can then appropriately estimate coefficients on variables such as military capabilities or democracy that factor into strategic utility equations. Signorino uses this method to estimate coefficients on various individual utility and probability terms that factor into the IIG. For this purpose, his method is preferred. In this case, because we are not estimating these coefficients, we do not need to use these methods.

[17] Many statistics texts discuss multinomial logit in more detail, including those by Hanushek and Jackson (1977) and Greene (1993).

coefficients and standard errors, simple t-tests on individual coefficients may not be adequate to establish whether some variable had an effect in determining outcomes. Coefficients in each equation give the effects of some variable *relative* to a base outcome. They suggest only whether the variable has a significant effect in distinguishing between the base category and the category in question, and selecting a different base category may *appear* to make different coefficients more or less significant. Individual t-tests cannot be relied upon as a sole means of assessing variable significance across the full model. Instead, two related techniques are relevant. First, a statistically significant coefficient in *any* equation suggests that a variable is important in differentiating between the base case and that equation. Thus, we should have confidence given any significant coefficient that the variable has *some* effect on outcomes. Second, we can use likelihood ratio tests to assess whether a variable has any effect on the overall model. Dropping any individual variable from a multinomial logit model results in a model that is nested within the original model. We can assess twice the difference in the log-likelihood ratios of the two models in a χ^2 test to assess whether or not a variable had a statistically significant effect on predicting overall outcomes. Following this logic, we use a series of block tests to assess the statistical significance of our independent variables across multiple outcomes.

Substantive effects are also difficult to interpret in multinomial logit. Neither the direction nor magnitude of the effect of a variable on outcomes can be interpreted by directly inspecting the multinomial logit coefficients. Rather, the marginal effects (both direction and magnitude) of a variable on any outcome can be calculated only given a knowledge of variable coefficients across *all* outcomes. The best way to interpret marginal effects in multinomial logit is to produce estimated probabilities of each outcome using the complete set of coefficients and various sets of independent variable values. Since the predicted probabilities from multinomial logit estimations for any single case must sum to 1, whenever the probability that a case will have one outcome rises, the probability of some other outcome(s) necessarily must fall. We apply the "method of recycled predictions" to generate these predicted values. Generally, the method of recycled predictions keeps actual independent variable values from actual cases, but sets the value of one particular variable of interest to a specific value for all cases to generate new predicted probabilities. We thus start with our initial set of data and the set of coefficients estimated in our model. We then change the value on the equilibrium variable of interest to a given value. We then use the coefficients from the model to create predicted probabilities for each of the outcome categories for each case. At this point, each case in the data has associated probabilities computed *as if* the case had the modified value of the independent variable of interest, with all other variables kept at their actual measured values.[18] We then compute the average probability across the dataset of different outcomes. This average probability reflects the probabilities that we *would have* observed if all cases had the modified value. Finally, we present the *relative risk* of observing different outcomes at different values of the equilibrium variables. Relative risk scores show the amount that the probability of an outcome changes (e.g., the probability of an acquiescence or a negotiation doubles or triples, for instance) without focusing on the absolute probability of non–status quo outcomes, which is quite small given the rarity of dispute behavior in international politics.

[18] This procedure is useful because of a disadvantage of a more typical method of analyzing predicted probabilities and changing by calculating the probability of some hypothetical case. Such a hypothetical case, typically created by setting variable values to their mean, may not reflect any case in the actual dataset. Moreover, analyzing probabilities across the full dataset takes into account the distribution of data on all variables.

d. Independent Variables

Expected Utility Data and IIG Equilibria. The independent variables used for testing the predictions of the IIG are dummy variables marking the game's equilibrium given utility values for a dyad. Ultimately the IIG under complete information conditions predicts that one of five outcomes could occur in equilibrium. These are an acquiescence by A (the initiator), acquiescence by B (the target), a negotiation, a war started by A, or the status quo. The other three outcomes are not predicted in equilibrium under the assumption of complete information. Since these are dummy variables and are theoretically mutually exclusive, only four can be included in the analysis at any time; we exclude the status quo as representing the baseline condition of peace from which we are interested in deviations.

Equilibrium predictions are made following the logical rules for how preferences combine to yield equilibria following the conditions specified by Bueno de Mesquita and Lalman (1992:71–90). The equilibria result from the way in which the sixteen theoretical utilities $U^i(SQ)$, $U^j(SQ)$, $U^i(Acq_i)$, $U^i(Acq_j)$, $U^j(Acq_i)$, $U^j(Acq_j)$, $U^i(Nego)$, $U^j(Nego)$, $U^i(Cap_i)$, $U^i(Cap_j)$, $U^j(Cap_i)$, $U^j(Cap_j)$, $U^i(War_j)$, $U^i(War_j)$, $U^j(War_i)$, and $U^j(War_j)$ interact.[19] The utilities for a state i are in turn defined by Bueno de Mesquita and Lalman (1992:47) as a function of eight input values. These include: a state's utility for the status quo $U^i(SQ)$; its utility for its most preferred international political position vis-à-vis the opponent $U^i(\Delta_i)$ (which it could impose in the case of victory); its utility for its least preferred international political position $U^i(\Delta_j)$ (which would be imposed on it in the case of loss); its subjective probability that it would win in a military confrontation P_i; the domestic political cost ϕ of using force rather than diplomacy to resolve differences; the magnitude of the advantage in costs α vs. τ of fighting away from home (on the target's soil) vs. being forced to fight on one's own soil; and the size γ of the cost of absorbing a first strike to which i gives in. These utilities are in turn defined and operationalized in Appendix 1 of *War and Reason.* In that appendix, states' utilities are in turn operationalized using risk attitudes and the similarity of foreign policy preferences, which is measured using the tau-b similarity score of their alliance portfolios. The subjective probability of winning in a military confrontation is estimated using the Correlates of War project's national capabilities index and estimates of the probability of intervention by third parties. Third party intervention probabilities are themselves estimated using the third parties' capabilities, foreign policy similarity toward i and j, and risk attitudes. The measurement of risk attitudes is detailed in Bueno de Mesquita (1985) as a function of alliance patterns and the hypothetical best and worst security/alliance situations in which a state might find itself.

Appendix 1 of *War and Reason* contains the detailed formulas for the full operationalizations of the terms needed to estimate utilities and in turn the final IIG equilibrium predictions (with two clarifications made below). However, since a number of data generation steps must be followed in sequence to create these estimates, we walk through this series of computational steps necessary to proceed from raw data to the final estimates of game equilibria:

- The individual components of national capabilities are assembled into the COW national capabilities index for every country-year (Singer, Bremer, and Stuckey, 1972).

[19] These terms using i and j are employed in chapter 2 of Bueno de Mesquita and Lalman, 1992, to designate theoretical utilities, while terms using A and B are developed in the Appendix to designate the empirically estimated utility values. In all cases, A corresponds to i and B corresponds to j.

- COW alliance data are used to calculate the tau-b score for each directed dyad in each year in the international system (Bueno de Mesquita, 1978, 1981).[20] Such scores are directed because the states included in computing tau-b scores depend on the region in which they are expected to interact (Bueno de Mesquita, 1981:94–97).
- Geographic location and contiguity data are used to create the distance between each pair of states, with changes over time due to major territorial changes taken into account.[21]
- Distance, tau-b, and capability data are combined to create an estimate of the expected utility of war for each directed dyad-year. The initial EU computation follows the methods in *The War Trap,* but does not adjust expected utility for risk attitude using *The War Trap*'s method. That is, the scores generated are the sum of the bilateral and multilateral expected utility components (equation 6, Bueno de Mesquita, 1981:59), but does not take the final step of introducing risk attitude.
- The initial expected utility values generated following the *War Trap* methods are used to produce estimates of states' risk attitudes for all years and with respect to each region in the system following the method of Bueno de Mesquita (1985). First, alliance data and expected utility data are used to generate actual (realized) national security portfolios for all state-years. Then, actual and hypothetical alliance data are used to generate hypothetical maximum and minimum national security portfolios for all state-years. Identifying hypothetical alliance patterns that minimize and maximize security involves a search over literally millions of possible alliances. In a region of twenty-one states with any of four alliance types possible with each other state, any one state has the possibility of allying in 4^{20} overall alliance patterns. Expected security under the hypothetical alliance portfolios is compared to the states' actual security to compute each state's annual risk attitude; the closer a state is to its security-maximizing alliance portfolio, the more risk-averse it is.[22]
- Tau-b scores and risk attitude scores are combined following the methods of *War and Reason* (equations A1.1 to A1.6, Bueno de Mesquita and Lalman, 1992:293–294) to produce estimates of states' utility for the status quo, their most preferred international political position vis-à-vis the opponent, and their least preferred international political position. These are the estimates of $U^A(SQ)$, $U^A(\Delta_A)$, and $U^A(\Delta_B)$ for each directed dyad-year A–B.

[20] Recent work by Signorino and Ritter (1999) has suggested an alternative to tau-b, namely, the "s" measure of similarity, theoretically superior to tau-b. Because we want to keep fundamental measures the same as Bueno de Mesquita and Lalman, we do not use s here.

[21] *War and Reason* did not incorporate distance discounting into its calculations because the empirical analysis was purely within Europe, where all geographic distances are relatively small. Since we are analyzing dyads worldwide, we reintroduce the distance discounting methods of *The War Trap.* We generalize and make uniform the method for computing state-to-state distance by taking the latitude and longitude of international cities and applying the "great circle" navigation distance formula to compute distance (Fitzpatrick and Modlin, 1986). For most dyads, we used the national capitals as the ends of the curve to compute distance. However, in the case of the U.S. and USSR, we used multiple cities as described in *The War Trap.* In addition, we considered countries that are contiguous on land to be 0 miles apart. We used the 1993 version of the COW contiguity dataset for this computation. We discount capabilities in each dyad ij so that (1) the capabilities of the challenger i are adjusted by the distance to the target j, and (2) the capabilities of third parties k are adjusted by the shorter of the distance to i or j. The states k are those states that might contribute capabilities to help i or j, and include states that are involved in the region of expected conflict for the directed dyad ij. We followed Bueno de Mesquita's definition of the relevant region for conflict (1981:97).

[22] We implemented a genetic algorithm (Holland, 1975; Goldberg, 1989) in our software to speed search over this space, but computation of risk scores remains the most time-consuming part of data generation. The generation of complete risk data from 1816 to 1984 required over six months on two continuously running PC 200 MHz Pentium Pro PCs. Users can take advantage of the data we generated without having to repeat this procedure.

- Tau-b scores, distance data, national capability data, and risk attitude scores (for states A, B, and third-party states within the relevant region of conflict) are combined following equations A1.7 to A1.10 of Bueno de Mesquita and Lalman, 1992:294–297, to produce estimates of the probability of success P^A for a state A in a military conflict against B, taking into account the likely behavior of possible interveners.
- The domestic cost term ϕ is measured (Bueno de Mesquita and Lalman, 1992:297) as $U^A(SQ)$. Following the actual method of constructing variables that Bueno de Mesquita and Lalman used (Bueno de Mesquita, personal communication), the other cost terms α, τ, γ are set to 1.0, because although these terms are defined theoretically and included as part of the expected utility equations, Bueno de Mesquita and Lalman had no way to measure them empirically. Without this assumption, complete measurements could not be constructed.[23]
- The utility measurements $U^A(SQ)$, $U^A(\Delta_A)$, and $U^A(\Delta_B)$, the probability of success P^A, and cost terms are combined following Table 2.2 of Bueno de Mesquita and Lalman, 1992:47, to produce estimates of the utility of each of the eight outcomes of the IIG for each directed dyad-year.
- The equilibrium outcome of the IIG played under complete and perfect information conditions is computed for each directed dyad-year using the two states' utilities for each of the eight IIG outcomes, following the logical conditions in Bueno de Mesquita and Lalman, 1992:72–92.[24] For every directed dyad-year, the sixteen final expected utility values corresponding to the utilities of states A and B are used to compute equilibria. Within the set of all dyads, the actual distribution of equilibrium predictions is such that approximately 12 percent of the cases have a status quo equilibrium prediction; 61 percent negotiation; 0.07 percent acquiescence by A; 0.5 percent acquiescence by B; and 29 percent a war initiated by A. Within politically relevant dyads, the distribution of equilibria is status quo 15 percent; negotiation 65 percent; acquiescence by A 0.1 percent; acquiescence by B 1 percent; war initiated by A 22 percent.[25]

[23] An additional implication of setting $\alpha = \tau$ (and so not measuring the magnitude of a first-strike advantage) is that the utility of War_A and War_B is always equal. The fact that α, τ, γ are set to 1.0 allows us to make unique equilibrium predictions for each dyad (with the exception mentioned in footnote 24), enabling us to avoid most of the problems, discussed by Signorino (1999), of dealing with unknown cost parameters.

[24] Because the utility of War_A and War_B is always equal (because $\alpha = \tau$, above), there remains one indeterminate situation in solving the game. Bueno de Mesquita and Lalman's Domestic Proposition 3.1, the "Basic War Theorem," specifies that logically a war will be started by state A if and only if the conditions $U^A(War_A) > U^A(Acq_A)$, $U^A(Cap_A) > U^A(War_B)$, $U^B(Cap_A) > U^B(Nego)$, and $U^B(War_A) > U^B(Acq_B)$ hold. However, with $U(War_A) = U(War_B)$, the first two conditions are incompatible, as they together imply $U^A(Cap_A) > U^A(War) > U^A(Acq_A)$, which violates the basic theoretical condition of the game that $U^A(Acq_A) > U^A(Cap_A)$. Because of this, the condition $U^A(Cap_A) > U^A(War_B)$ is omitted for purposes of generating the war equilibrium (see Bueno de Mesquita and Lalman, 1992:76n). However, with this condition omitted, it is possible for a directed dyad-year to satisfy the conditions for both the War A and Status Quo equilibria; in our data, 22,987 out of 805,439 (2.8%) directed dyads (and 4,321 cases out of 122,712 politically relevant dyads, 3.5%) are indeterminate. In the analyses reported here, we treat these indeterminate cases as falling into the War_A equilibrium category; the results do not change significantly if we treat these cases as having a status quo equilibrium prediction instead.

[25] It is interesting to note a few features of this real-world distribution that will clearly have an effect on its later fit to outcome data. It is clear from these results and the actual frequency of conflictual interactions in the international system that the status quo is underpredicted by the IIG and associated operationalizations, while negotiation and war are overpredicted. With this clue, future work could focus on the logical conditions and associated operationalizations of these equilibria in particular. For instance, it may be that measures of the utility of the status quo relative to other outcomes are underestimated in the operationalizations used by Bueno de Mesquita and Lalman, or that measures of the possible gains from conflict are overestimated by differences in foreign policy portfolios (tau-b scores), despite discounting capabilities for distance. It could also be that the assumption that the war cost terms α, τ, γ equal 1 (made for empirical estimation purposes) is problematic, and in fact the costs of war loom much larger than the model gives those costs credit for. Alternatively, if we believe that

We updated a number of elements in carrying out these procedures relative to the computations conducted by Bueno de Mesquita and others in previous works. In generating the data for our analyses, we used updated COW alliance and national capability data. We interpolated data on British energy consumption when it was missing for the 1816–1850 period (these data were often missing in this period). We used the newest version (v2.1) of the COW MID data to identify disputes and dispute outcomes, and added distance discounting to the methods of *War and Reason.* Finally, we used an improved and more accurate algorithm (compared to Bueno de Mesquita, 1985) to generate risk attitude scores. Of course, as Bueno de Mesquita and Lalman are careful to point out in their own work (1992:299), these operationalizations are still likely to have measurement error in them that will make the fit between the game equilibrium and actual outcomes imperfect at best. Assuming the measurement error is random, however, the bias this induces results in an attenuation of the estimated values for the parameters estimated for the equilibrium conditions. Absent measurement error, we would expect the true relationship to be substantively stronger than that presented below.

To perform these tasks, we developed a software program titled *EUGene,* for Expected Utility Generation and data management software (Bennett and Stam, 1998a, 2000c). This program follows the methods of Bueno de Mesquita and Lalman (1992), uses the most up-to-date data available, and brings the multiple steps necessary to create expected utility data (initially published across a number of journals over a long time period) together in a single, easily accessible program. Until recently, to carry out these computations one needed a fast mainframe computer. The computationally intensive nature of these computations (most notably for risk attitude) limited Bueno de Mesquita and Lalman's testing in *War and Reason.* More broadly, it is because not all of the essential intermediate variables have been available that broad testing of the IIG has been impossible until now. A short technical description of the program can be found in the Appendix.[26]

In sum, for each directed dyad-year from 1816–1984, we created a prediction of which of the eight game outcomes *should* result, assuming that (1) the actors are playing the game, (2) our data are accurate, and (3) the actors have complete information. Under these conditions, because of logical restrictions imposed by the game, only five outcomes (the status quo, acquiescence by state A, acquiescence by state B, negotiation, and war) are possible in equilibria. We are then left with a prediction for each directed dyad-year of one of five equilibrium outcomes, which we represent as four dummy variables.

Controls. One criticism of the original tests of the IIG in Bueno de Mesquita and Lalman, 1992, is that the tests of the model's predictions were conducted in most cases with few controls. It is important to include controls to check the robustness of the IIG results (they turn out to be fairly robust), because variables

the IIG measures are adequate, it could be our measures of conflict frequency that are at fault, if in fact the MID data we commonly employ in analysis underestimate the occurrence of conflict because they measure only high-level conflicts. Thinking about the underestimation of negotiation, it is useful to note that the IIG predictions are made assuming complete information. Given real-world uncertainty, the negotiation outcome is likely to be affected more than other outcomes because there are many possible conditions that may be satisfied to lead to a negotiation outcome, and any violations would lead to other outcomes. (If we add a simple control variable marking uncertainty to our analysis, as in Bueno de Mesquita and Lalman 1992:77, 216, our empirical findings about the negotiation equilibrium do not change, however.) These puzzles about the predictions of the game remain for future analysis.

[26] Our software (which is available as freeware, including source code) allows these computations to be carried out on a PC, allows our methods to be inspected, and allows replication or modification by other analysts. Without this software, generation of the data used here would not have been possible.

drawn from other theories may be correlated with the equilibrium predictions from the IIG, and in fact explained variance being ascribed to the IIG may not be related to the game but to coincidence with other theories. As a result, in some of our analyses we include a set of control variables drawn from other theories of conflict that represent the major theoretical foci of recent international conflict studies, including regime type, power, interests, and opportunity. In some cases, these variables represent sub-components of the overall IIG equilibria (in particular, the separate tau-b score and balance of forces). These individual components are combined in a different, nonlinear fashion within the IIG to create equilibrium predictions, and so there is no necessary reason for the components to be highly correlated with the dummy equilibrium prediction variables. Even where the IIG purports to have an explanation that encompasses other explanations (for example, *War and Reason* discusses how the democratic peace may fit within the context of the game and rational preferences), it is important to know whether other variables add explanatory power to our understanding of conflict, and to what extent they "take variance away" from the IIG and its associated measures.[27] We explore the sensitivity of our results concerning the IIG to the inclusion of the following auxiliary variables:

Balance of Forces. The notion that relative power affects conflict behavior is one of the oldest and most basic notions in world politics. It is also one of the most hotly contested. Balance of power theory holds that when power is equal between states, conflict will be less likely than when it is unequal. The power preponderance perspective holds just the opposite. Beyond balance of power, work in rational deterrence theory (e.g., Huth, 1988) suggests that the greater the military advantage for a state, the more likely it will issue a military challenge or escalate a conflict. We use the composite national capabilities score from the Correlates of War (COW) capability data as our measure of military capability (Singer, Bremer, and Stuckey, 1972). We create a ratio of the potential initiator's capabilities to the total capabilities of the dyad. The final variable ranges from 0 (when the initiator is weak, and has no capabilities relative to the target) to 1.0 (when the initiator possesses 100% of the dyad's capabilities).

Tau-b Score. A key variable that contributes to expected utility calculations is the similarity in foreign policy preferences between the potential initiator and potential target, measured in terms of tau-b scores between alliance portfolios. We include this score as a separate variable to see if similarity has an effect on outcomes beyond that captured in the game equilibria. The variable ranges from -1 to $+1$, with positive values representing increasingly similar alliance portfolios.

Contiguity. While distance in general is captured in the game equilibria through our discounting of military capabilities by distance, we also include a dummy variable marking when the potential initiator and target are contiguous on land. In most cases, contiguity removes all transportation barriers for military forces between states, and serves as an indicator of conflict salience. We would expect contiguity to play a role in determining both the initiation and escalation of disputes between states.

Democracy. The large body of literature on the so-called democratic peace derives from the simple empirical observation that no two democracies have ever fought a war against each other, an observation that dates back over two decades (Small

[27] While unlikely, it is also possible that the inclusion of control variables could strengthen the findings for the IIG.

and Singer, 1976). Doyle (1983) developed an explanation of this empirical observation based on Kant's notion of perpetual peace among liberal states. Russett (1990) and Maoz and Russett (1993) extended this work with a focus on shared norms of compromise and the presence of restraining democratic institutions that prevent democracies from fighting one another. Broader arguments have also been developed that encompass the democratic peace, for instance, the notion that it is regime similarity rather than simply shared democracy that helps states avoid conflict (e.g., Raknerud and Hegre, 1997). These arguments are different from those posited by Bueno de Mesquita and Lalman, which concern the domestic costs of conflict to democratic leaders, and beliefs about dovishness. If democracy has effects on conflict that lie outside the IIG, our inclusion of the democracy variables will help to capture it.

We build on Oneal and Russett's most recent measures of regime type (Oneal and Russett, 1997). We begin with individual regime type scores for each state in a dyad. These are measured as the state's democracy score, minus the autocracy score, in the given year from the Polity III dataset (Jaggers and Gurr, 1995). This index ranges from -10, indicating states with low democracy and high autocracy, to a $+10$, indicating states with the opposite. Since our study is both dyadic and directed, we want to be able to capture both the monadic effects of the initiator and target democracy on conflict, and the interaction. We therefore include four variables. We start with separate indicators of initiator and target democracy to pick up linear trends about regime type and conflict. We add an indicator of the difference in the regime scores of the two sides, allowing us to examine whether similar regimes are more or less likely to become involved in conflict. Fourth, we interact the initiator's regime score with the difference to incorporate effects of democracies (or autocracies) initiating and escalating disputes differentially against particular other types of states. Finally, to capture purely dyadic propositions about the democratic peace, we convert the regime type scores to be strictly positive (so they range from 0 to 20) and multiply them together, resulting in a variable that is high only when both states have highly democratic regimes.

Peace Years Spline. One criticism of studies of conflict using pooled dyadic time-series has been that they fail to take into account time dependence within dyads over time (Beck et al., 1998). We account for cross-dyad differences in our data by correcting for clustering by dyad. We correct for time dependence by following the Beck et al. method of including a set of four spline variables that take into account the time (number of "peace years") that has passed in a dyad since a prior dispute.

Results: Universal Tests of Expected Utility Theory

In this section, we present the results of our tests of Bueno de Mesquita and Lalman's rendering of expected utility theory in the IIG, using data developed by the *EUGene* software. We begin our analysis by looking at a model that only includes the IIG equilibrium variables, examining both the subset of politically relevant dyads and the set of all dyads. The coefficient estimates we obtained by analyzing the relationship between the IIG equilibria and behavior within politically relevant dyads are given in Table 3, while those we obtained when conducting our analysis on the set of all dyads are given in Table 4. As we stated above, it is sometimes difficult to tell from a set of individual coefficient estimates in multinomial logit whether or not a variable is really having a "significant" effect on outcomes. As a result, we conducted a series of block tests to look at the effect of the equilibrium variables on the explanatory power of the model as a whole; these tests are presented in Table 5. We present estimates of the

TABLE 3. Multinomial Logit Coefficients, Politically Relevant Dyads

Outcome:	*Acquiescence by B*		*Negotiation*		*Capitulation by A*		*Capitulation by B*		*Use of Force*	
Variable	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*
Equilibrium: Acquiescence by A	1.192	1.040	2.010	1.063	−39.655	—	0.711	1.030	−43.131	—
Equilibrium: Acquiescence by B	0.505	0.500	−44.691	—	−41.613	—	−1.585	0.992	0.394	0.829
Equilibrium: Negotiation	0.113	0.217	−0.219	0.321	1.824	1.018	0.049	0.166	0.318	0.196
Equilibrium: War by A	0.467	0.218	−0.184	0.351	2.433	1.033	0.450	0.198	0.807	0.220
Constant	−0.603	0.207	−6.847	0.289	−9.555	1.000	−5.547	0.154	−5.917	0.182

n = 122712; log likelihood = −10437.324

TABLE 4. Multinomial Logit Coefficients, All Dyads

Outcome:	*Acquiescence by B*		*Negotiation*		*Capitulation by A*		*Capitulation by B*		*Use of Force*	
Variable	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*
Equilibrium: Acquiescence by A	1.324	1.026	2.200	1.048	−40.616	—	0.718	1.015	−43.027	—
Equilibrium: Acquiescence by B	0.961	0.493	−43.607	—	−42.615	—	−0.562	0.717	0.855	0.839
Equilibrium: Negotiation	0.036	0.211	−0.310	0.319	0.579	0.606	−0.156	0.151	0.180	0.190
Equilibrium: War by A	0.058	0.212	−0.644	0.346	0.882	0.626	0.012	0.180	0.355	0.207
Constant	−7.619	0.203	−8.495	0.290	−10.104	0.577	−7.013	0.140	−7.514	0.177

n = 805439; log likelihood = −15216.963

TABLE 5. Block Tests of Statistical Significance, All and Politically Relevant Dyads, Model without Controls

	Politically Relevant Dyads *(n = 122,712)*			*All Dyads* *(n = 805,439)*		
Variable Removed	*df*	*Log-Likelihood*	*Probability*	*df*	*Log-Likelihood*	*Probability*
(Null Model, constant only)	—	−10480.33	—	—	−15231.98	—
(Full Model, All Equilibrium Variables)	16	−10437.32	0.003 (vs. null model)	16	−15216.96	0.118 (vs. null model)
Acq_A Equilibrium	3	−10439.47	0.23 (vs. full model)	3	−15219.35	0.19 (vs. full model)
Acq_B Equilibrium	3	−10441.65	0.034	3	−15221.16	0.04
Negotiation Equilibrium	5	−10442.72	0.055	5	−15219.35	0.44
War_A Equilibrium	5	−10462.28	<0.001	5	−15222.34	0.056

TABLE 6. Predicted Outcome Probabilities: Relative Risk Associated with Equilibrium Conditions, Model without Controls

	Politically Relevant Dyads						*All Dyads*					
Equilibrium:	*SQ*	*SQ*	*AcqA*	*AcqB*	*Nego.*	*WarA*	*SQ*	*SQ*	*AcqA*	*AcqB*	*Nego.*	*WarA*
	Probability	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Probability*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*
Outcome:												
Status Quo	0.98997	1.00	0.99	1.00	1.00	0.99	0.99782	1.00	0.997	1.00	1.00	1.00
Acquiescence by B	0.00239	1.00	3.25	1.66	0.11	1.58	0.00049	1.00	3.75	2.61	1.04	1.06
Negotiation	0.00105	1.00	7.37	—	0.80	0.83	0.00020	1.00	8.99	—	0.73	0.53
Capitulation by A	0.00007	1.00	—	—	6.19	11.30	0.00004	1.00	—	—	1.78	2.42
Capitulation by B	0.00386	1.00	2.01	0.21	1.05	1.56	0.00090	1.00	2.05	0.57	0.86	1.01
Joint Force or War	0.00267	1.00	—	1.49	1.37	2.23	0.00054	1.00	—	2.35	1.20	1.43

Relative risk indicates the relative effect of changing values of the independent variables on outcome probabilities.

Risk greater than 1.0 indicates a higher risk of observing the outcome than under the status quo equilibrium; risk below 1.0 indicates lower risk.

effects of the equilibrium variables on outcomes (probabilities and relative risks) in Table 6.

The subset of politically relevant dyads is where we believe the IIG equilibria should fit best, and is where we have previously found the IIG to successfully predict conflict outcomes (Bennett and Stam, 2000a). An examination of t-ratios in Table 3 suggests that equilibrium predictions of acquiescence by A (the initiator), acquiescence by B (the target), and war have statistically significant effects on the actual occurrence of conflict outcomes. The negotiation equilibrium does not appear to have a clear influence on differentiating between the status quo and other outcome categories. When we examine the block likelihood ratio tests in Table 5, however, the negotiation equilibrium does appear to add explanatory power to the model, suggesting that this prediction may be differentiating between other categories of behavior. In Table 5 it appears that removing any of the equilibrium predictions except for acquiescence by A results in a significant decrease in model fit. Taken together, these results suggest that the equilibria are significantly correlated with state behavior, at least within the set of politically relevant dyads.

It takes a more thorough examination of the direction of the effects of the equilibrium predictions on actual behavior to reach conclusions about how well the IIG helps us understand conflict, however. In Table 6, we present the relative risk of observing various outcomes given different equilibrium predictions. In all cases, we begin with the specification that the relative risk of observing an outcome is 1.0 when we predict that the status quo will be the equilibrium. The other risk ratios give the predicted risk, relative to the status quo, when a different equilibrium is predicted. Ratios greater than 1 indicate increased risk, while ratios less than 1 indicate diminution of risk. Table 6 then suggests, for example, that the mutual use of force is approximately 2.2 times more likely to occur in dyads where the IIG equilibrium is war than in dyads where the equilibrium is status quo. Similarly, when the predicted equilibrium was for an acquiescence by B, the relative risk of observing a capitulation by B was 0.20, or an 80 percent *reduction* in the risk of observing that outcome.

At the most general level, an inspection of the risk ratios in Table 6 shows that with three exceptions, all of the non–status quo equilibria do in fact correlate with non–status quo behavior. In 13 out of 16 instances, equilibrium predictions other than the status quo do correspond to a more frequent observation of non–status quo outcomes. At a more specific level, we can look at the correspondence between specific equilibrium predictions and outcomes. When the IIG predicts acquiescence by the conflict initiator, we see more acquiescence by the *target,* more negotiation, and more capitulation. We are unable to estimate how much acquiescence by the initiator we actually see, because we simply do not have enough cases of that outcome. We are also unable to estimate the effect on war outcomes because a joint use of force or war never occurred in the data when we predicted an acquiescence by A. Nevertheless, given this prediction of a low-level non–status quo outcome, we do see non–status quo outcomes more often, and do not see escalation to the highest level of conflict within the dataset. Looking at predicting acquiescence by the target, we see significantly more acquiescence by the target, no negotiation or capitulation by the initiator, less capitulation by the target, and more escalation to high levels of conflict.

Examining the equilibrium prediction of negotiation relative to the status quo, three outcomes occur with significantly increased or decreased frequency. First, we actually find a *lower* probability of negotiation when we predict negotiation (counter to what we would expect). Second, we see a higher probability of capitulation by the initiator and higher probability of force than when the status quo is predicted. Finally, when the IIG predicts war, we see significantly less negotiation, significantly more capitulation, and significantly more uses of force and war outcomes (a 2.2 times increase).

Of the equilibria where we both have predictions and can measure the precisely corresponding outcome, then, only in the case of negotiation does the expectation not square with reality. Within politically relevant dyads, rather than a subset of European dyads, we actually find that the negotiation equilibrium is associated with somewhat *less* negotiation. The drop in the probability of negotiated settlements is not as dramatic as the increases in the probability of other outcomes given different equilibria, however.

Many other specific relationships in Table 6 fit our expectations about the effects of the expected utility equilibria. For instance, when war is the equilibrium, we see large increases in the probability of war, capitulation (backing down under a threat of force), and acquiescence. We also see a decrease in the probability of negotiation. Thus, when war is expected, we do often get war, but we also see states backing down unilaterally, perhaps in recognition that war is a likely result if they continue. This is consistent with signaling explanations of war initiation. The fact that we do not see negotiation when we expect war is also reasonable, as we might expect states in this situation to either give up under pressure or hold out until we get the expected war, but might not expect them to sit down and resolve disputes in a more amicable fashion.

Our next analysis is of the population of all directed dyad-years. Rational choice theorists do not claim that their theory should be limited in time or space, and so might expect that the same relationships found in politically relevant dyads should hold regardless of the setting. However, by way of counterargument, if there is significant variation in either the quality of the data, the actors' preferences, or the game the actors play across time or region, then we would not expect to see a robust relationship between equilibrium predictions and actual behavior in the set of all dyads. For instance, it may be that the alliance-based utility measures are more accurate among politically relevant dyads simply because these are the states that are most active in forming alliances, and who look to alliances as reliable indicators of intentions. Given less variance in our utility measures and common and conflicting interests that are not reflected in alliances, our measures may be less reliable when we look at "politically irrelevant" dyads.

First, looking simply at the statistical significance of the overall model, there is doubt that the IIG equilibria overall gives us any improvement in fit over the null model. However, individual coefficients and standard errors in Table 4 and the block tests in Table 5 suggest that some of the equilibria marginally help improve model fit. From the coefficients in Table 4, the war equilibrium predicts differences in outcomes from the status quo to other categories. It is also likely that the negotiation and acquiescence by A equilibria correlate with deviations from the status quo ($p = .07$, .10, and .06 in various relationships). There is no clear relationship between acquiescence by B and the status quo (or other outcomes). The block tests tell a slightly different story about overall power, suggesting that the acquiescence by B equilibrium and the war equilibrium are contributing to overall model fit while the other equilibria do not. Overall, it is less clear how well the equilibria fit with actual outcomes outside the politically relevant subset of dyads.

When we turn to the relative effects of the equilibrium variables on predicting outcomes among the set of all dyads (Table 6), the pattern of relative risks is quite similar to that within the set of politically relevant dyads. In no case does any equilibrium correspond to more risk within the politically relevant dyads and less (or vice versa) within the population of all dyads. In 13 out of 16 instances again, equilibrium predictions other than the status quo do correspond to our observing non–status quo outcomes. With a few exceptions, the magnitude of the effects is also quite similar. As before, the effects of the negotiation equilibrium are an exception. Given a negotiation equilibrium, the risks of a capitulation by A drops from a sixfold increase among politically relevant dyads to only a 1.8 times increase among all dyads. We note a similar drop in the estimated effect of the war equilibrium on capitulation by A. For the majority of the equilibria, the relative risk associated with the equilibria is slightly attenuated when compared to the results for the politically relevant sample. Since we are dealing with relative risks, this change is not because non–status quo outcomes are less frequent among the set of all dyads, but instead suggests that the equilibria are in general not predicting as well among all dyads. One notable change of particular substantive interest to point out from Table 6 is the estimated effect of the war equilibrium on the actual use of force. Within the set of politically relevant dyads, when the IIG equilibrium is war, we observe more than a doubling of the actual risk of war. Among all dyads, the risk of war increases by only 40 percent. The equilibrium prediction appears simply not to have as much explanatory power among the set of all dyads. This is sensible if we believe that dyads outside the politically relevant sample are not playing the IIG, or that our measurements do not apply as well. It also suggests that we need to rethink the automatic assumption that the IIG will apply and predict equally well everywhere.[28]

Overall, in our models without controls for other explanations, the IIG equilibria appear to generally predict actual non–status quo outcomes. However, the correspondence between the equilibria and outcome behavior is clearer and stronger within politically relevant dyads. This remains a much wider set of cases for applying the theory than has been used previously, of course. In many ways, finding any relationships in our data is remarkable given that we have analyzed an extremely skewed sample of data consisting mostly of non-events, and where some would argue that the noise in such a large number of cases would wash out any findings. These findings suggest that predicting behavior using the methods and measures developed by Bueno de Mesquita and Lalman (1992) have broader applicability than the small sample of cases they presented. What remains to be seen, however, and what we turn to next, is whether these findings will still hold up after we include a set of controls for other important explanations of international behavior.

Controlling for Alternative Explanations

Because the initial equilibrium-only model may be mis-specified as a model of conflict, Tables 7 through 10 present the coefficients and relative risk scores obtained for analyses run with a fuller model. Because of missing data on the independent variables, in these analyses we lose about 12 percent of our cases within the politically relevant dyads, and about 18 percent of our cases within the set all dyads. Individual coefficients in this model now suggest that only the negotiation equilibrium and war equilibrium variables are helping to differentiate non–status quo outcomes from the status quo among either the set of politically relevant or all dyads.[29] When we examine block significance tests, however, we find a slightly more complicated picture. It is very clear that even after we have added other controls, the overall set of four equilibrium variables does significantly improve the fit of the model ($p < 0.001$ among politically relevant dyads, and $p < 0.003$ among all dyads). However, there is uncertainty about which specific equilibrium variables are providing that improvement, as drop-

[28] In analysis we do not report here we examined just the subset of "politically irrelevant" dyads and found that the equilibria simply do not predict well at all. This is due in large part to there being very few non–status quo outcomes among these dyads, certainly a smaller proportion than even among all dyads. When we analyze politically irrelevant dyads, for instance, we cannot estimate the effects of either acquiescence equilibrium, for instance, or the effects of any equilibrium on capitulation by A.

[29] The apparently large coefficients on the acquiescence by A and acquiescence by B equilibria in fact indicate that our statistical software cannot obtain good estimates.

Table 7. Multinomial Logit Coefficients, Model with Controls, Politically Relevant Dyads

Outcome:	*Acquiescence by B*		*Negotiation*		*Capitulation by A*		*Capitulation by B*		*Use of Force*	
Variable	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*
Equilibrium: Acquiescence by A	1.155	0.986	0.880	1.014	−39.572	—	−43.522	—	−43.319	—
Equilibrium: Acquiescence by B	0.693	0.565	−43.719	0.389	−40.692	1.148	−1.500	1.009	0.068	0.872
Equilibrium: Negotiation	0.204	0.244	−0.637	0.318	1.806	1.028	−0.032	0.183	0.062	0.217
Equilibrium: War by A	0.439	0.241	−0.460	0.342	2.053	1.037	0.338	0.200	0.431	0.224
Balance of Forces	1.051	0.175	0.580	0.282	1.067	0.376	0.264	0.157	0.303	0.190
Tau-b Score	−0.071	0.166	0.350	0.267	−0.388	0.515	−0.028	0.212	−0.051	0.190
Democracy Score, State 1	0.102	0.047	0.132	0.068	−0.105	0.107	0.006	0.039	0.046	0.047
Democracy Score, State 2	0.075	0.041	0.192	0.058	−0.001	0.087	0.068	0.029	0.027	0.030
Difference in Dem. Scores	−0.021	0.026	−0.100	0.040	0.081	0.068	0.000	0.024	0.001	0.026
State 1 Democracy × Difference	0.000	0.003	0.002	0.004	0.007	0.005	0.000	0.002	−0.002	0.003
Democracy 1 × Democracy 2	−0.008	0.003	−0.017	0.005	0.005	0.008	−0.004	0.003	−0.004	0.003
Contiguous on Land	1.281	0.163	1.826	0.249	1.491	0.308	0.765	0.163	1.850	0.195
Peace Years	−0.278	0.042	−0.463	0.079	−0.334	0.101	−0.465	0.041	−0.552	0.046
Peace Years Spline 1	−0.002	0.001	−0.004	0.001	−0.001	0.002	−0.004	0.001	−0.005	0.001
Peace Years Spline 2	0.001	0.000	0.002	0.001	0.000	0.001	0.002	0.000	0.002	0.000
Peace Years Spline 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Constant	−4.761	0.614	−3.325	0.907	−9.947	1.724	−3.644	0.513	−4.125	0.633

n = 108094; log likelihood = −7690.889

TABLE 8. Multinomial Logit Coefficients, Model with Controls, All Dyads

Outcome:	*Acquiescence by B*		*Negotiation*		*Capitulation by A*		*Capitulation by B*		*Use of Force*	
Variable	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*	*Coef.*	*Std. Err*
Equilibrium: Acquiescence by A	0.943	0.973	0.808	1.005	−27.841	0.883	−29.216	0.332	−28.216	0.369
Equilibrium: Acquiescence by B	0.801	0.571	−28.579	0.397	−27.091	0.850	−0.781	0.765	0.122	0.891
Equilibrium: Negotiation	0.044	0.241	−0.684	0.316	0.581	0.616	−0.213	0.174	−0.034	0.220
Equilibrium: War by A	0.278	0.242	−0.644	0.346	0.823	0.637	0.034	0.187	0.272	0.229
Balance of Forces	1.418	0.215	0.620	0.311	1.321	0.425	0.081	0.193	0.488	0.218
Tau-b Score	−0.173	0.178	0.226	0.265	−0.688	0.625	−0.127	0.218	−0.145	0.193
Democracy Score, State 1	0.144	0.046	0.133	0.067	−0.125	0.108	0.002	0.040	0.043	0.047
Democracy Score, State 2	0.078	0.039	0.177	0.056	−0.050	0.089	0.043	0.030	0.025	0.029
Difference in Dem. Scores	−0.040	0.026	−0.094	0.040	0.112	0.071	0.023	0.025	0.003	0.026
State 1 Democracy × Difference	−0.002	0.003	0.002	0.004	0.006	0.005	−0.001	0.002	−0.002	0.003
Democracy 1 × Democracy 2	−0.009	0.003	−0.016	0.005	0.008	0.008	−0.002	0.003	−0.004	0.003
Contiguous on Land	3.036	0.153	3.699	0.238	3.403	0.279	2.545	0.151	3.662	0.179
Peace Years	−0.307	0.039	−0.453	0.075	−0.340	0.096	−0.505	0.041	−0.569	0.043
Peace Years Spline 1	−0.002	0.001	−0.004	0.001	−0.001	0.002	−0.004	0.001	−0.005	0.001
Peace Years Spline 2	0.001	0.000	0.002	0.001	0.000	0.001	0.002	0.000	0.002	0.000
Peace Years Spline 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Constant	−6.226	0.607	−5.247	0.917	−11.379	1.688	−5.425	0.532	−5.951	0.631

n = 667179; log likelihood = −10445.457

Table 9. Block Tests of Statistical Significance, All and Politically Relevant Dyads, Model with Controls

	Politically Relevant Dyads (n = 108,094)			*All Dyads (n = 667,179)*		
Variable Removed	*df*	*Log-Likelihood*	*Probability*	*df*	*Log-Likelihood*	*Probability*
(Null Model, constant only)	—	−9027.62	—	—	−12996.11	—
(Full Model, All Equilibrium Variables)	75	−7690.89	0.003 (vs. null model)	75	−10445.46	<0.001 (vs. null model)
All 4 equilibrium variables	15	−7715.43	<0.001 (vs. full model)	15	−10462.70	0.003 (vs. full model)
Acq_A Equilibrium	2	−7692.73	0.06	2	−10447.42	0.15
Acq_B Equilibrium	5	−7695.58	0.095	3	−10449.29	0.06
Negotiation Equilibrium	5	−7696.13	0.07	5	−10449.14	0.20
War_A Equilibrium	5	−7702.22	<0.001	5	−10450.25	0.07

Table 10. Predicted Outcome Probabilities: Relative Risk Associated with Equilibrium Conditions, Model with Controls

	Politically Relevant Dyads						*All Dyads*					
Equilibrium:	*SQ*	*SQ*	*AcqA*	*AcqB*	*Nego.*	*WarA*	*SQ*	*SQ*	*AcqA*	*AcqB*	*Nego.*	*WarA*
	Probability	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Probability*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*	*Relative Risk*
Outcome:												
Status Quo	0.98883	1.00	1.00	1.00	1.00	1.00	0.99759	1.00	1.00	1.00	1.00	1.00
Acquiescence by B	0.00219	1.00	3.18	2.02	1.23	1.53	0.00047	1.00	2.59	2.23	1.05	1.31
Negotiation	0.00154	1.00	2.43	—	0.53	0.62	0.00029	1.00	2.28	—	0.51	0.52
Capitulation by A	0.00008	1.00	—	—	6.08	7.61	0.00004	1.00	—	—	1.80	2.26
Capitulation by B	0.00404	1.00	—	0.23	0.97	1.38	0.00096	1.00	—	0.46	0.81	1.03
Joint Force or War	0.00332	1.00	—	1.09	1.06	1.50	0.00064	1.00	—	1.14	0.98	1.30

Relative risk indicates the relative effect of changing values of the independent variables on outcome probabilities.

Risk greater than 1.0 indicates a higher risk of observing the outcome than under the status quo equilibrium; risk below 1.0 indicates lower risk.

ping individual equilibrium variables does not have the same clear effect on the model's log-likelihood. This is likely the result of collinearity between individual equilibrium variables and some controls in the model.

Examining the relative risk scores, adding the set of additional control variables has a strikingly similar effect on the predicted risk ratios in both samples of dyads. With few exceptions, the addition of additional substantive variables reduces the increase in risk associated with equilibrium outcome variables, but does not eliminate them. For instance, in the politically relevant dyad subset, the relative risk associated with the war equilibrium drops from a factor of 2.2 to 1.5, while amongst all dyads, the risk drops from 1.4 to 1.3. A few of the apparent reductions in risk are more dramatic, such as the drop in the risk of observing negotiation given a prediction of A's acquiescence from 7.4 to 3.2. In only one instance in each table have the coefficients changed so much that we shift from predicting an increase in risk (in the absence of controls) to predicting a decrease in risk once we include substantive control variables. This occurs in the all-dyads sample, with the negotiation equilibrium predicting on observing the target's capitulation, and with the negotiation equilibrium predicting the use of force (in both cases the effects flip-flop near 0). This suggests that the effects of predicting equilibria on outcomes is weakened, but not eliminated, by including additional controls. In general, even with the inclusion of additional controls, predicting non–status quo equilibria corresponds to observing more non–status quo outcomes in most instances in both sets of cases, with or without controls. The biggest anomalous finding from prior analyses remains: when the game equilibrium is negotiation, we actually see relatively *fewer* negotiated outcomes. Nevertheless, the clearest and arguably most important finding also remains intact, namely, that there is a 30 to 50 percent increase in the odds of observing escalation to the "deepest" conflictual stages in the game tree when war is the equilibrium prediction rather than the status quo.

It is important to note that this effect is not due primarily to the inclusion of variables that are directly included in calculations leading to equilibrium predictions (we include the balance of capabilities and tau-b score separately in these controlled analyses). Further analysis (tables not presented here) reveals that while the separate tau-b and balance of forces measures do have significant explanatory power (from assessing a decrease in the log-likelihood), dropping them does not systematically or dramatically change the relative risk of the IIG equilibria. For example, when we include the separate tau-b measure in the analysis of all dyads, we find that when the IIG predicts an acquiescence by A, the relative risk of observing an acquiescence by B is 2.58, and the relative risk of observing negotiation is 2.28. When we drop the separate measure, the first risk score drops to a relative risk of 2.18, but the second simultaneously increases to a relative risk of 2.82. Our findings about the power of the IIG equilibria are generally quite robust. Overall, the pattern of predicted probabilities supports the fit of expected utility theory predictions to dispute initiation and escalation down the IIG game tree. Importantly, the results presented here confirm this with many more cases, across multiple regions, in additional years beyond those in prior analyses, and once we include controls for additional explanations about international conflict and for individual model sub-components. However, we also find that the predictions of the theory do not correspond as well to actual outcomes when we expand our set of cases too far, namely, to include the many so-called politically irrelevant dyads in the international system. This is an important finding to bear in mind as the analysis and application of expected utility theories is expanded in the future. This may also prove to be an important element to be taken into account in developing future expected utility models and operationalizations to be applied in a variety of international settings.

Conclusions

In this paper, we have completed the most in-depth test to date of whether or not the expectations deduced from Bueno de Mesquita and Lalman's IIG successfully predict militarized conflict and other types of behavior. We did this after developing software to generate expected utility data following the methods of Bueno de Mesquita and Lalman (1992). Overall, our results support the prior findings of Bueno de Mesquita and Lalman that the game-theoretic predictions of the IIG and the empirical measures developed to test them successfully predict behavior. We also found, though, that while the game equilibria do correlate with international conflict behavior, they are more powerful predictors within a restricted subset of data (politically relevant dyads) than expected utility theorists might have expected. In addition, not all equilibria find the same level of support. These important caveats must be kept in mind when assessing rational choice theory as it has been applied in international relations.

Ex post, we can construct plausible explanations for these findings. For instance, decision-makers in most nonpolitically relevant dyads simply may not be playing a game anything like the IIG. Leaders where there is no realistic possibility of applying force may not even consider the possibility of making a demand backed up by a threat. Absent an army, it is hard to imagine how Costa Rica's strategic calculations closely approximate those represented in the IIG, for example. Other possible explanations lying outside the application of the IIG exist as well, such as those based in variable construction. Regardless of the theoretical explanation, our empirical findings raise a red flag for the straightforward application of the IIG to the international system writ large, and must be explored further.

The existence of a complete dataset of expected utility information and updated risk scores for all dyads between 1816 and 1984 opens up numerous possibilities for additional research. The fit of the IIG equilibria within subsets of data can be explored (see, e.g., Bennett and Stam, 1998b), as can its differential fit to "joiners" into MIDs vs. the initiators of MIDs. Further exploration can be made of the interaction of uncertainty in the international system with expected utility predictions about dispute behavior. With these data now readily available, we can also see how expected utility predictions of war mesh with alternative explanations for international conflict. We turn to these questions in other projects.

Appendix: Expected Utility Generating Software (*EUGene*)

EUGene (Expected Utility Generation and data management program) combines the 1992 methodology of *War and Reason* with an easy-to-use program to calculate expected utility values. The program generates up-to-date risk scores, raw utility values, and predicts the equilibrium of the *War and Reason* IIG for all dyad-years from 1816 to 1984. As data are made available, the software will reflect the most recent version of the COW, MID, and Polity data. *EUGene* uses several datasets as raw inputs, including national capability data, alliance data, country independence dates, geographical contiguity, and location data. The program then computes the COW national capabilities index, tau-b scores for all dyads, risk attitudes of all states for all regions, and states' utility for the eight possible outcomes of the IIG. In a final step, *EUGene* uses the states' utility scores to predict the outcome expected in equilibrium for each directed dyad-year in the interstate system. The program provides users with options for modifying these calculations; for instance, users can modify the capability data used as input and regenerate utility predictions for sensitivity analysis. *EUGene* is the first program to implement Bueno de Mesquita and Lalman's (1992) methodology to generate data for the full set of dyads from 1816 to 1984.[30] We are able to conduct a broad test of Bueno de Mesquita and Lalman's expected utility theory only because we developed this program.

EUGene runs under Microsoft Windows 95 (or higher) or Windows NT (version 4.0 or higher) on IBM-compatible PCs with at least 16 MB of memory. The program was written in the Borland, Inc. *Delphi* programming environment. The program is copyrighted, but is available as a free download from our website at *http://www.eugenesoftware.org.* The downloadable program includes the main program executable file (about 700K), complete (60-page) documentation of the program, *EUGene*'s source code, and the complete expected utility data described in this paper (approximately 50 MB of data). In addition to providing instructions for how to use the program, the program documentation details the computations involved in making expected utility calculations, key algorithms used in the program, and describes a series of tests conducted to confirm that the program was generating data as we expected.

In addition to generating risk and expected utility data, *EUGene* also serves as a useful tool to simplify the merging and creation of datasets in international relations. EUGene will create datasets with different units of analysis (country-year, dyad-year, dispute-dyad-year), for different subsets of cases (e.g., all dyads, politically relevant dyads, or major power dyads), and including a variety of variables such as Polity III democracy scores (Jaggers and Gurr, 1995), tau-b scores, risk attitude data, contiguity data, and Correlates of War Militarized Interstate Dispute data (users can currently select from a set of over 60 variables to be included in the output data file). All of these data are converted to a common dyadic format and merged for output. *EUGene* is designed to make a number of cumbersome tasks associated with building international relations datasets much easier by automating a variety of tasks that must be performed to integrate several of the most important other datasets in international relations. A number of other datasets are downloaded with *EUGene* for merging to create output data files with multiple variables. For users who wish to merge data themselves, *EUGene* can also be used to generate a simple set of case identification data (country codes and years) for all dyads, politically relevant dyads, and so on. As it creates datasets, *EUGene* also creates command files that can be used to read the dataset into SPSS, Stata, or LIMDEP.

References

BECK, N. (1996) Reporting Heteroskedasticity Consistent Standard Errors. *The Political Methodologist* **7**(spring):4–6.

BECK, N, J. KATZ, AND R. TUCKER (1998) Taking Time Seriously: Time-Series–Cross-Section Analysis with a Binary Dependent Variable. *American Journal of Political Science* **42**(Oct.):1260–1288.

BENNETT, D. S., AND A. C. STAM (1998a) *EUGene: Expected Utility Generation and Data Management Program.* Website: http://www.eugenesoftware.org.

BENNETT, D. S., AND A. C. STAM (1998b) Is Instrumental Rationality a Universal Phenomenon? Paper presented at the Annual Meeting of the Midwest Political Science Association, April, Chicago.

BENNETT, D. S., AND A. STAM (2000a) Cross-Validation of Bueno de Mesquita and Lalman's International Interaction Game. *British Journal of Political Science,* forthcoming.

[30] A prior software program, *Tolstoy* (Horn, 1990), developed under Bueno de Mesquita's auspices, did not employ the most recent methods of Bueno de Mesquita and Lalman (1992). *Tolstoy* allowed data for only a limited set of dyads to be generated, and contained errors (e.g., some risk values output by the program fell outside the theoretically allowed range of −1 to +1). As *EUGene* was developed, we communicated frequently with Bueno de Mesquita, who provided us with feedback and clarifications and specified a number of details of auxiliary assumptions made during the development of the data used in *War and Reason.* The complete source code and details of our software implementation are available for download and inspection with the program.

Bennett, D. S., and A. C. Stam (2000b) *Research Design and Estimator Choices in the Analysis of Interstate Dyads: When Decisions Matter. Journal of Conflict Resolution*, forthcoming.

Bennett, D. S., and A. C. Stam (2000c) EUGene: A Conceptual Manual. *International Interactions* **26**:179–204.

Bueno de Mesquita, B. (1978) Systemic Polarization and the Occurrence and Duration of War. *Journal of Conflict Resolution* **22**:241–267.

Bueno de Mesquita, B. (1981) *The War Trap.* New Haven, CT: Yale University Press.

Bueno de Mesquita, B. (1985) The War Trap Revisited. *American Political Science Review* **79**:156–177.

Bueno de Mesquita, B., and D. Lalman (1992) *War and Reason.* New Haven, CT: Yale University Press.

Bueno de Mesquita, B., D. Newman, and A. Rabushka (1985) *Forecasting Political Events.* New Haven, CT: Yale University Press.

Doyle, M. (1983) Kant, Liberal Legacies, and Foreign Affairs, Parts 1 and 2. *Philosophy and Public Affairs* **12**:205–234, 323–353.

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

Fitzpatrick, G. L., and M. J. Modlin (1986) *Direct-Line Distances,* U.S. ed. Metuchen, NJ: Scarecrow Press.

Goldberg, D. E. (1989) *Genetic Algorithms in Search, Optimization, & Machine Learning.* New York: Addison-Wesley.

Green, D. P., and I. Shapiro (1994) *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science.* New Haven, CT: Yale University Press.

Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems.* Ann Arbor: University of Michigan Press.

Horn, M. (1990) *Tolstoy: The Computer Program of War and Peace.* Computer program. University of Rochester.

Huber, P. J. (1967) "The Behavior of Maximum Likelihood Estimates under Non-standard Conditions." In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability.* Berkeley and Los Angeles: University of California Press.

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

Jaggers, K., and T. R. Gurr (1995) Tracking Democracy's Third Wave with the Polity III Data. *Journal of Peace Research* **32**:469–482.

Jervis, R., R. N. Lebow, and J. Stein (1985) *Psychology and Deterrence.* Baltimore, MD: Johns Hopkins University Press.

Jones, D. M., S. A. Bremer, and J. D. Singer (1996) Militarized Interstate Disputes, 1816–1992: Rationale, Coding Rules, and Empirical Patterns. *Conflict Management and Peace Science* **15**:162–213.

Kaufmann, C. D. (1994) Out of the Lab and into the Archives: A Method for Testing Psychological Explanations of Political Decision Making. *International Studies Quarterly* **38**:557–586.

Lebow, R. N. (1981) *Between Peace and War.* Baltimore, MD: Johns Hopkins University Press.

Maoz, Z., and B. Russett (1993) Normative and Structural Causes of the Democratic Peace. *American Political Science Review* **87**:624–638.

Oneal, J. R., and B. M. Russett (1997) The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985. *International Studies Quarterly* **41**(June):267–294.

Raknerud, A., and H. Hegre (1997) The Hazard of War: Reassessing the Evidence for the Democratic Peace. *Journal of Peace Research* **34**(4):385–404.

Russett, B. (1990) *Controlling the Sword: The Democratic Governance of Foreign Policy.* Cambridge, MA: Harvard University Press.

Signorino, C. S. (1999) Strategic Interaction and the Statistical Analysis of International Conflict. *American Political Science Review* **93**(June):279–297.

Signorino, C. S., and J. M. Ritter (1999) Tau-b or Not Tau-b: Measuring the Similarity of Foreign Policy Positions. *International Studies Quarterly* **43**(Mar.):115–144.

Simowitz, R., and B. L. Price (1990) The Expected Utility Theory of Conflict: Measuring Theoretical Progress. *American Political Science Review* **84**:439–460.

Singer, J. D., S. Bremer, and J. Stuckey (1972) "Capability Distribution, Uncertainty, and Major Power War, 1820–1965." In *Peace, War and Numbers,* edited by B. Russett. Beverly Hills, CA: Sage.

Small, M., and J. D. Singer (1976) The War-Proneness of Democratic Regimes, 1816–1965. *Jerusalem Journal of International Relations* **1**:50–69.

Smith, A. (1999) Testing Theories Involving Strategic Choice: The Example of Crisis Escalation. *American Journal of Political Science* **43**(Oct.):1254–1283.

White, H. (1978) A Heteroskedasticity Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica* **46**:817–838.