2 COMPARATIVE HYPOTHESIS TESTING AND SOME LIMITS TO KNOWLEDGE

it is a test of true theories not only to account for but to predict phenomena.
—William Whewell, 1840

To model the changes in the risk of war associated with the various models and arguments in our analysis we employ a variant of the maximum likelihood model of inference described by King (1989). Following this approach, and using a series of nested statistical models, we can assess whether the independent variables in our models improve our ability to predict systematically relative change in the risk of conflict between nation-states. For each argument or conjecture we test, we derive one or more independent variables predicted by the original author to correlate with the dependent variable (international conflict and escalation thereof). We then assess (1) whether the independent variable(s) associated with each explanation makes a statistically significant contribution to the overall likelihood of the wars that occur in our data and (2) the relative explanatory power of each variable on a comparable scale. If a variable marking a prediction derived from some "theory" does not contribute to the overall likelihood of the statistical model, or makes little substantive difference in the relative risk of conflict onset, we must question whether that particular argument helps us to understand international conflict in a systematic way compared to being a description of an idiosyncratic event.

We are able to examine and evaluate indicators drawn from multiple levels of analysis by focusing on the directed dyad-year as our unit of analysis. A directed dyad-year is a pair of states in a given year, observed from the perspective of one of the two states. This distinguishes the identity of a potential conflict's initiator and target. For example,
China-Japan in 1990 is one observation, while Japan-China in 1990 is another. In either observation, a conflict could occur. Because each dyad-year contains two states, this choice allows us to incorporate variables such as those from balance of power approaches that require information on two states. It allows us to include unit-level information arguably associated with conflict initiation, such as a potential initiator’s regime type or rate of democratization. Finally, we can also examine the power of system-level arguments using this unit of analysis. If the international system in one period is more conflict prone than in another, this will be reflected in a higher probability of multiple states choosing to initiate conflict.¹

This process of comparative hypothesis testing helps to address several important problems with existing tests of empirical international politics propositions. The first is the problem of overlapping predictions and controls. It is important to include multiple explanatory factors in a single model even if those different explanations appear to be independent. Commonly, the independent variables marking separate arguments correlate to some degree, particularly since many of the variables constructed by theorists draw their inspiration from the same underlying sets of cases. If this is the case, then controlling for the presence of the other variables is critical to minimize the possibility of omitted variable bias, claims of theoretical independence notwithstanding. For example, Bueno de Mesquita and Lalman (1992) present their International Interaction Game (IIG) as an essentially complete data-generating process, but this is so only in the context of the assumptions they need to develop their theory. Even if two competing explanations are mutually exclusive in that they begin with different assumptions (and few theories of international politics are truly so exclusionary or well developed), we must include appropriate predictive variables from both in the same model to properly assess which empirical conjecture better fits the data while controlling for the predictions of the other.

If we only analyze variables or explanations in isolation, we also tend to overlook the epistemological limitations that multicollinearity can pose. For example, NATO emerges as an effective security organization at about the same time that the United States and the USSR both develop secure second-strike nuclear capabilities, the conditions needed to follow the doctrine of Mutually Assured Destruction (MAD). While it is true that no large-scale open war occurred between the United States and the USSR, it is impossible for us to tell whether NATO or MAD (or some other collinear explanation) is responsible for the absence of war using this type of statistical analysis. If we find that the
empirical predictions of an argument do not allow us to predict unique events due to severe multicollinearity, we then have a fundamental problem, an inherent limit to knowledge associated with the method used here.

The second inferential problem that comparative hypothesis testing can help address is the possibility that multiple factors might simultaneously affect the likelihood of states initiating an international conflict. States may be inclined to choose war over peace at different times due to the confluence of multiple “weak factors” rather than as a result of a single provocation or set of incentives. Alternatively, it may be that one set of factors drives international conflict at one time or location and that a different set of factors may be more relevant at some other place or point in time. This would imply that there are regionally or temporally distinct paths to war. While some scholars have long argued that multiple factors affect international politics (Waltz 1959), debate continues over what level of analysis, approach, or theory provides the “best” depiction of international politics. A more fruitful question to ask is when, where, and how much each source of conflict influences international conflict behavior. It may be that international power concerns generally dominate domestic political issues, or that international system characteristics trump dyadic factors, or that both play an important role, perhaps under different circumstances that can be specified ex ante. However, we cannot judge the relative risks associated with each of these perspectives by focusing on one explanation at a time. Our method of combining multiple variables in a single equation may also help us understand the relatively weak fit of existing models of international conflict cast at the dyad-year level. If each of the many causal descriptions of international conflict explains separately but simultaneously a piece of the empirical world, then it is only by combining the key predictor variables from those arguments into a single statistical model that we can maximize our predictive power.

LIMITS TO KNOWLEDGE:
EPISTEMOLOGICAL PERSPECTIVES

Our perspective here is generally a behavioral one, following in the tradition of those for whom observable behavior is the ultimate outcome begging explanation. From the perspective of the traditional behavioralists, we would hope to be able to explain 100 percent of the observed variation in our dependent variable. While we draw from this
behaviorist well, we nevertheless believe that there are several reasons why any statistical model will not come close to approaching perfect predictive accuracy. In this section, we discuss some recent arguments that suggest limits on our ability to predict systematically international conflict.

Typical behavioral research seeking to develop a general explanation for some phenomenon (that is, accounting for variance in the dependent variable) implicitly starts with the ubiquitous regression equation

\[ Y = \alpha + \beta X + e, \]  

(2.1)

where \( Y \) is the dependent variable, \( \alpha \) is a fixed intercept, \( \beta \) represents a set of parameters applied to a vector of explanatory variables \( X \), and \( e \) represents errors assumed to be stochastic. The \( X \) vector includes the set of variables whose associated effects analysts are interested in studying. In the context of international politics, these might include state power, domestic political institutions, international trade, or several other measurable variables. More precisely, if we are working with a cross-sectional time-series dyadic data set (data that tracks dyads across both space and time) where \( i \) and \( j \) represent different states (and so \( ij \) represent a particular directed dyad) and \( t \) represents time (\( ijt \) is the dyad \( ij \) at time \( t \)), our regression equation becomes

\[ Y_{ijt} = \alpha + \beta X_{ijt} + e. \]  

(2.2)

The standard analytic approach focuses attention on the signs and parameter estimates or coefficients of \( \beta \), seeking to minimize \( e \) by including the “right” set of independent variables in \( X \). We must correctly specify this set of independent variables in order to operationalize the critical aspects of a given theory or conjecture in order to measure the concept(s) in question with as little error as possible and to account for the interactive effects present in strategic situations (Signorino 1999, 2000; Lewis and Schultz 2001). Much of the statistical work on international politics essentially ignores the single fixed intercept \( \alpha \), assuming it theoretically uninteresting. Nor do scholars typically explore variations in \( \beta \) or \( e \) over either time (\( t \)) or location (\( ij \)) (see Green, Kim, and Yoon 2001; Cederman 2001 is a notable exception). Fortunately, however, econometricians have generalized this equation and developed classes of models appropriate for data that consist of a pool of cross-unit, cross-time data. A general form of equation (2.2) is the following:
Equation (2.3) shows both $\alpha$ and $\beta$ potentially varying by unit and across time. Econometricians have developed models specifically designed to deal with variations in the intercept ($\alpha$) across units and variations in coefficients ($\beta$) across units and time. Recently, scholars in international politics have also focused attention on the error term $e_{ijt}$ and suggested that due to the strategic nature of political decision making there may be an upper limit to how much we can explain using measurable variables included in either $\alpha$ or $\beta$ (Gartzke 1999). Next, we explore each of these modeling perspectives and approaches before turning to the arguments and hypotheses we examine for possible inclusion in our vector of variables, $X$.

**Smoothed Line**

**Intercept Variation:**

*The Changing Nature of Politics over Time and Region*

Statistical studies of international conflict typically assume that a single intercept ($\alpha$) is adequate. The estimate of $\alpha$ reflects the baseline probability, frequency, or rate of conflict. However, if $\alpha$ varies so that different regions, countries, or dyads have different baseline risks of conflict not accounted for in the vector of variables $X$, then estimating a single $\alpha$ is inappropriate and may bias our estimate of $\beta$. Similarly, if $\alpha$ varies over time, as would be the case if a dyad’s baseline or underlying level of conflict varied from one period to another, so that the intercept in one year differs from that in another, then a single $\alpha$ is again inappropriate. The intercept $\alpha$ may even vary by both time and space, so that the baseline level of conflict differs by dyad (or region) and time period. We take up this problem in detail in chapter 6.

The idea that a dyad’s intrinsic level of conflict might change or evolve over time is particularly attractive to constructivists. Wendt (1999) argues, for example, that a dyad’s underlying tendency to resort to war is a function, in part anyway, of some mutual socially constructed identity that is subject to change over time. In the language of our statistical model, if $\alpha$ (the baseline level of conflict) in our model is not fixed but instead is specific to some aggregation unit (e.g., region, dyad, or year), then a model with a single intercept risks omitted variable bias. If this unaccounted variance in $\alpha$ is correlated with some of the variables in $X$, the vector of parameters $\beta_{ijt}$, then the parameters we estimate (the effects on the rate of conflict associated with the
independent variables of interest) will likely be biased as well. In essence, if different countries or dyads have different baseline conflict propensities not accounted for in the vector of variables $X$, the covariance that should be associated solely with unit differences may then be attributed to (some of) the coefficients in $\beta$. The econometric solution is to provide for the estimation of an intercept for each of the dyads in the study. The class of fixed-effects estimators was developed to deal efficiently with this problem; if $\alpha$ varies by unit in ways that cannot be accounted for by the vector of variables $X$, then a fixed-effects model may be appropriate (Green, Kim, and Yoon 2001; Bennett and Stam 2000a).

In practice, only a few studies in international politics have used fixed-effects models. The most prominent recent application is by Green, Kim, and Yoon (2001), who argue that the problem of overaggregated intercepts is common in dyad-year studies of international conflict. The authors find important differences in their results when they retest existing models of conflict using a fixed-effects estimator. However, this approach is not a cure-all. Beck and Katz (2001) and King and Zeng (2001) point out some significant problems with fixed-effects estimators. They argue that the solution may be worse than the problem. For instance, if we specify a different intercept for each dyad we cannot estimate the effect of independent variables that do not vary over time even though the cross-sectional variation associated with them may be of interest (such as geographic contiguity, which varies spatially but is time invariant for most dyads). Perhaps, more importantly, with fixed-effects models we cannot use information gleaned from those dyads where there is no variation on the dependent variable. This implies we can learn nothing about what leads to peace by studying the dyads that have never gone to war. In a fixed-effects model, independent variables that do not change over time add nothing to the model’s overall fit to the data, since they correlate perfectly with the unit intercept and drop out of the estimation. Dyads for which the dependent variable is constant (e.g., those that never have conflicts) similarly contribute nothing to the likelihood function, and they too drop from the analysis. Importantly, the statistical necessity of dropping these variables and observations from fixed-effects analyses has nothing to do with whether they actually influence the onset of war (Beck and Katz 2001).

If we truly believe that variables such as contiguity that change only across space and not over time are irrelevant (and that they are not correlated with other variables of more theoretical interest) and that dyads without conflict really provide us with no information about the process that leads to conflict in those dyads that suffer from it, then a fixed-effects estimation might be appropriate. However, we do not believe that
such variables and dyads are without information and thus agree with those who believe that fixed-effects estimation may be overkill. The problem with mistakenly assuming a single fixed intercept \( \alpha \) is again that the relevant and omitted cross-unit effect, \( \alpha_{ijt} \), becomes associated with the other variables in \( \beta \). We can solve the problem without resorting to the fixed-effects model by properly specifying the causal factors that lead \( \alpha_{ijt} \) to vary for some different dyads or groups of dyads. That is, by specifying the factors that cause different dyads to have a different baseline level of conflict. The solution to this thorny empirical problem is more and better theory development rather than brute force statistical fixes. We believe that the better alternative to fixed-effects modeling is to consider carefully whether we should expect different units to have different intercepts for theoretical reasons.³ By properly specifying variables to be included in \( X \) that capture this unit variation, we will essentially move conflict covariance from \( \alpha \) to an individual parameter \( \beta \).

The key element of this process is then to specify properly at what levels of aggregation we expect to find the sources of international conflict. Several possibilities suggest themselves as units to investigate. We might speculate that there are effects that differ by particular years (if some year is more or less dangerous for all states than another year, for instance, in the year after the conclusion of a large war) (Smith and Stam 2002; Werner 1999). Alternatively, groupings of years into eras might make sense if, for instance, we believed the cold war era to be more or less dangerous than the interwar period. Spatially, groupings by state or state type may be called for (for instance, if individual states are more dangerous by virtue of regime type or economic growth). Finally, grouping by region may be reasonable if we believe that states in particular regions engage in different patterns of international behavior by virtue of culture, shared colonial experience, learned interaction patterns, or peculiar regional politics. If these factors (spatial or temporal specific effects) are substantively important, but omitted from the analysis, the ignored differences could bias our parameter estimates. Regardless of the nature of this type of bias, the solution, in the end, is again more and better theory than exists today.

In chapter 6, we will explore whether regional and temporal variations in the fit of rational choice models of international conflict are significant. As we show there, variation in \( \beta \) over space and time proves important. An interesting implication of this is that the distinction between international politics and comparative politics—two traditionally separate subfields in political science—blurs tremendously; the remaining distinctions between the subfields essentially lie in one’s normative choice of interesting or important dependent variables.
A second type of misspecification suggested in equation (2.3) occurs if the coefficients in $\beta$ vary over units and time. If a conjecture “works” or a hypothesis holds up for only some dyads, or for only some time periods, and not for others, then assuming that a single $\beta$ is appropriate for all members of $X$ is again a potential source of bias, inefficiency, or both. If we believe that some hypothesis should hold only in a subset of cases, or at some particular time, then we can easily modify the variables, thereby controlling for this source of bias. For instance, we could account for an argument linking the effects of bipolar international systems to the behavior of just the great powers by including an interaction term of the dummy variable “bipolarity” with “major power.” This would produce a third variable capturing the effects associated just with major powers in a period of bipolarity. Situations where the effects of variables change smoothly over time are more difficult but possible to deal with. This might be the case, for instance, if we believed that the balance of capabilities effectively mattered less than institutional democracy as time passes (Wendt 1999; Cederman 2001).

If we have specific theoretical arguments specifying how coefficients should change over time, then one relatively efficient solution is to include particular interactive variables to estimate the expected effect, for instance “time since last dispute” $\times$ “balance of forces.” If we believe that coefficients vary in relation to continuous time, then time-varying-parameter models are another way to address this problem (Cederman 2001). We investigate this problem in chapter 6, where we use time-interactive dummies to model how the effects of various expected utility equilibria conditions evolve over time.

Uncertainty, Rational Expectations, and the Limits to Knowledge

Recent work by Fearon and others suggests that there may be an upper bound to the predictability of conflict due to the strategic nature of the decisions leading to international war. The argument is perhaps made most clearly by Fearon (1995; see also Blainey 1988), who argues that it is critical for rationalist theories of international politics to consider why wars ever occur, given that fighting is an inefficient conflict resolution or bargaining mechanism. In any situation where two (or more) sides disagree over the distribution of some stake, rational parties
should seek the most efficient (least costly) solution to the disagreement. However, violent and expensive contests such as wars impose large costs on the participants in the process of achieving some settlement. If rational, and assuming the two sides gain no utility from expressive behavior, both parties to such an expensive contest should prefer ex ante to settle the disagreement at the final terms without suffering the costs of the contest.

This poses a puzzle for theorizing about war: since states expect war to be costly, why do they fight before reaching a mutually agreeable settlement? Fearon argues that uncertainty about the war’s likely outcome is perhaps the most general explanation of why conflict occurs in the international system. If the parties to a disagreement shared accurate predictions about the costs and outcome of a contest, they would likely settle before the contest occurred. According to this logic, without uncertainty, we would have no costly contests (Fearon also points to potential problems of commitment and divisibility but focuses on the role of private information). Seeing such costly contests, uncertainty is likely a critical factor leading to an increased likelihood of war. Uncertainty in this perspective occurs in the form of private information held by actors and beliefs about factors such as military capabilities and resolve. As long as we cannot measure private information, we cannot directly test the theory, nor can we hope to be able to accurately predict with certainty which situations will lead to war and which will not.

One implication of this argument is that uncertainty is an important variable that should be included in our empirical models of conflict if we can measure its nature or magnitude. We could rewrite our statistical equation again, splitting our error term into two parts, one being uncertainty as discussed previously and one being the more conventional stochastic error with a mean of zero:

\[ Y_{ijt} = \alpha_{ijt} + \beta_{ijt}X_{ijt} + U_{ijt} + \varepsilon_{ijt}. \] (2.4)

In equation (2.4), \( U_{ijt} \) represents rationalist arguments about the systematic role of uncertainty as it affects our dependent variable, while \( \varepsilon_{ijt} \) includes truly stochastic or random unobservable effects, assumed to have a mean of zero. Uncertainty may have some distribution different from the distribution of the \( \varepsilon_{ijt} \) error term. Models that focus on information and signaling are really focusing on the \( U \) term as distinct from \( \varepsilon \).

Gartzke (2000) develops in detail the logical implications of this representation for what we can and cannot know about the initiation of interstate conflict. If we are unable to eliminate \( U \) (which by definition
we cannot), then there is an upper bound to the predictability of conflict. Gartzke argues persuasively that we may never be able to predict international conflicts with high levels of certainty. If Fearon and others are on the right track by emphasizing strategic uncertainty as the source of most conflict, the best we may be able to do is to identify situations where the likelihood of conflict occurring is a fifty-fifty proposition. Scholars face a theoretically imposed limit to how much we can understand about international conflict. As Gartzke (2000) provocatively puts it, the wars that actually occur do so only “in the error term”—not in the parameters we estimate because of the immeasurable influence of uncertainty. Again, this theoretical limit occurs because the private information that drives uncertainty is by definition immeasurable. In turn, if a critical factor in a model is immeasurable, it will necessarily cause the systematic variance in our model to fit at less than 100 percent. We might be able to predict at a higher than fifty-fifty rate either if we as analysts know what the leaders themselves do not (but this strikes us as implausible) or if leaders behave irrationally in systematic ways.

It is important to note that this necessity of imperfect fit does not emerge from the more conventional sources of error in the term $e$ in our initial model. In the naive behavioral approach using statistical analysis, the assumption is that if we could properly specify our theoretical model (that is, include all relevant theoretically specified variables), and if we could measure our concepts without error, then we should be able to achieve a 100 percent success rate in prediction (reflected in, for instance, an $R^2$ of 1.0). Gartzke’s argument suggests that, given rational decision makers, even if scholars specify everything in their model exactly right, their rate of correct predictions could not exceed 50 percent. It also leads to a slightly different interpretation of the effects associated with the other conjectures and their indicators found in $\beta$. Following Fearon and Gartzke’s logic, the other variables in our model then correlate with the situations where informational uncertainty is most relevant, and hence the situations where conflict is most likely to reach the upper bound of predictability.

We may well be able to identify dyads, time periods, countries, or other spatial and temporal aggregations of actors that are prone to conflict with some frequency. That is, we may be able to identify the times and places where the risk of conflict is roughly fifty-fifty compared to the international system’s overall risk of war of approximately $1/400$ dyad-years. But because the outbreak of any individual war or conflict is driven by uncertainty, and consistent with Riker’s (1980) gloomy expectations noted in chapter 1, we will never be able to make an accurate point prediction about the outbreak of any single war or crisis.
The nature of this theorized upper limit to prediction remains to be more completely developed. This limit frames the expectations we should have for any successful theory development and testing efforts in at least two important ways. First, it further strengthens our earlier argument about the probabilistic nature of theories and conjectures in international politics, although for a different reason (the nature of uncertainty and information rather than human “randomness” or misspecification). Second, it suggests that being able to identify situations where conflict is a fifty-fifty proposition may be the best we can do using observable ex ante data. As a result, we should temper our expectations for the performance of all statistical models of strategic behavior. While at times it may appear that empirical predictions are weak in highly strategic settings such as those leading to international conflict, in fact, the measures developed over the past forty years may be performing extraordinarily well once we understand the upper bounds to what these models can explain and what they cannot. Of course, the existence of this predictive limit hinges on the rationalist assumptions that international actors are engaged in bargaining and are concerned with avoiding ex post inefficiency.

It is important to note that this limit to knowledge in predicting conflict is not a problem associated solely with statistical (or, more generally, large-\(n\)) approaches. Rather, these limits exist for all approaches to understanding conflict where the aim is prediction rather than retrospective understanding. Since the source of the problem lies in uncertainty and private information, no methodology that relies on evidence can solve it. In fact, of the available testing approaches, statistical tools provide the most leverage on the problem. Over several iterations, we can understand tendencies and probabilities that allow us to approach the limit of predictive understanding. Assuming we are only able to predict wars with probabilistic accuracy, these predictions are reliable only in samples large enough to reduce random error to acceptable levels. While archival methods will continue to reveal ex post private information that decision makers held when they made the decisions leading to war (something the approach here cannot capture), they cannot predict future conflicts beyond the limits we have discussed, which affect all research.

COMMON QUESTIONS ABOUT COMPARATIVE HYPOTHESIS TESTING

Throughout this chapter, we have given our perspective on comparative hypothesis testing. Along the way, we have also addressed several possible
problems with this approach. Here, we bring together some additional possible concerns and attempt to address them more comprehensively.

Can You Test an Explanation of War with Just a Single Dependent Variable?

We test the arguments and conjectures presented later by comparing their predictions to the evidence. Our research design focuses on predictions and evidence in only one area of state behavior, international war. Some might argue that conflict theories might have important implications for a variety of areas other than or in addition to conflict initiation. For instance, balance of power advocates have used the concept of power balances to make predictions about alliance formation (Walt 1987), behavior under different system structures (Waltz 1979), the national imperative for expansion (Schweller 1996), and how violent conflict correlates with situations of power balance or imbalance (Mearsheimer 2001). These authors correctly point out that a theory is more than a prediction of a single empirical relationship. Rather, a theory is a set of closely related set of assumptions and logical arguments designed to illuminate a causal process. To test a theory we must assess its predictions across a number of events; for instance, if we see states’ actions following the patterns predicted by the theory and we observe relationships that the theory predicts in multiple areas, then we should be relatively more likely to accept the veracity of the theory.

Our concern here, however, starts from the opposite direction, namely, the puzzle of explaining violent international conflict and assessing the value of the arguments and conjectures that purport to be able to predict the relative risk of international conflict. From the behaviorist perspective, we do not want theories for the sake of having theory per se. Rather, we want theories that help us to understand and to predict future empirical patterns of behavior. In our case, we have narrowed our focus to the empirics of international war. If the evidence about when conflict is more or less likely does not support the key hypothesis from some theory, or if we do not find consistent support for some empirical conjecture about the relative likelihood of war, this raises important questions about the value of “international relations theory.” If an explanation of international conflict cannot predict conflict, either its likelihood or severity, then what good is it? Our concern is with the scientific nuances of prediction, not the aesthetics of retrospection. Some argue that balance of power models do not explain
much of the observable interstate violence and instead only suggest what kinds of coalitions will emerge (Schweller 1996; Levi 2002). This take on balance of power theory is not of interest to us here, as it does not predict the behavior we are seeking to explain and only explains possible prior steps in the conflict process.

What about Relationships between Theories?

Because theories are more than single predictions of behavior but instead posit a set of underlying linkages, some theories are logically compatible while others are not. For instance, balance of power and power transition arguments make opposite predictions about whether conditions of power equality increase or decrease the chances of conflict. Power transition theorists add further that it is not just equality, but equality plus the presence of a dynamic transition and dissatisfaction, that drives conflict (Lemke 2002). If we include variables from each argument in an additive statistical model, the contradictory causal logic of these two explanations makes it impossible for both sets of variables simultaneously to affect conflict behavior in the direction adherents expect. Whether it is due to weak operationalization or possible logical flaws in a theory, if our results suggest that two logically opposed arguments or propositions both work, it would be an indication that we must carefully examine the models, their assumptions, and the nature of the operational measures.

What Does Using an Additive Statistical Model Imply?

Our tests in chapter 4 primarily exploit an additive specification in a multinomial logit model. In practice, few if any theories of international conflict are comprehensive enough to provide a description of a complete data-generating process, which logically would exclude the possible validity of other arguments. As we noted earlier, most “theories” of conflict are actually not really theories at all but rather empirically based descriptions of some factor believed to influence the likelihood of interstate conflict. These so-called theories typically do not provide any explicit consideration regarding which other factors might also warrant simultaneous consideration or exclusion. By focusing on individual factors, authors implicitly assert that “this single causal factor matters and others do not,” or else they unwittingly assume away the problem of omitted variable bias in the tests at hand.
The story we try to convey throughout this book is that most theories or conjectures on the initiation and escalation of conflict simply are capturing isolated examples of the multiple factors that influence policymakers’ decision making. We know that many factors influence decision making in the real world; leaders are constantly subject to the push and pull of competing interest groups often with mutually exclusive preferences. Taken as a group, our models of international conflict reflect this understanding of the real world. However, individual tests of a single or a few factors in isolation do not. As Clarke (2001, 730–31) points out, we are testing a “supermodel” of international conflict that he also refers to as “artificial” and atheoretic. We believe that the models need not be, however, given the plausible assumption that many forces simultaneously influence leaders’ decisions and therefore the risk of conflict.

Can We Test a Theory with a Single Independent Variable?

Previously we discussed testing theories, hunches, and conjectures using a single dependent variable (the level of militarized conflict). It is equally important to consider whether we can test a “theory” using a single independent variable. Many models of international conflict specify simple and direct relationships between a single variable and conflict and as such are not really theories but rather hypotheses or conjectures that we can test in straightforward fashion. For instance, some long-cycle explanations of conflict suggest that particular global economic conditions are favorable to conflict. There is a single variable, “global economic prosperity,” that Goldstein (1988) argues should influence the dependent variable “international war.” In these cases, obviously, we need to only include a measure of the single relevant concept to be able to assess the empirical validity of the argument. Other models posit more complicated, interdependent relationships between any numbers of concepts that do not readily lend themselves to simplistic testing schemes. For instance, realists commonly argue that the anarchic character of the international system affects the nature and behavior of interstate alliances. The number and scale of alliances in some conceptions define polarity, which in turn affects the relatively likelihood of conflict. Given our concern here—predicting conflict—it is appropriate to focus on the final relationship in this process—the relationship between polarity and conflict. If there is no empirically verifiable relationship between polarity and conflict, then the concept of “polarity” does not help us to predict the onset of war and we must question whether and how
neorealism theory is useful to us (from the perspective of trying to understand better the onset of interstate disputes and war). In this example, if we find that alliances do affect conflict but polarity does not, then it would suggest that some parts of the realist causal argument are flawed.

**What about Conditionality?**

Some models or approaches predict that several variables should affect conflict, in some instances conditionally. By “conditionally,” we simply mean that the argument in question suggests that variables will have particular effects only in the presence of other conditions. For instance, power transition models predict that equal power between states contributes to an increased risk of conflict. But conflict should emerge only when a rough bilateral equality of power combines with a shift in the two states’ relative power, marking a transition from one state being in the lead to its suddenly being behind, simultaneously combined with the overtaking state being dissatisfied with the nature of the rules that constitute the international system. We can easily test propositions that specify such conditionality simply by building and including appropriate interactive variables. For the case of power transition models, such an interaction term could simply be “equal power” times “rapid power movement” times “satisfaction.” If there is either unequal power or no movement, the variable will have a value near zero, whereas only if power is equal and there is a transition will the variable have a value of one. Of course, if we only included the variables for “power” and “movement” separately, we would not properly capture the conditionality aspect of the power transition logic. As a result, the inclusion of appropriate interactions is critical for proper testing of many arguments and hypotheses.

A similar approach can account for arguments that focus on “context.” Signorino’s (1999) point that the inclusion of “component” utility variables may be inappropriate in the presence of strategic interaction is correct but easily remedied by the inclusion of terms that do capture strategic interactions such as point predictions of various equilibria.

Finally, for some arguments or conjectures, single variables quite accurately represent or capture the intuition of the argument. For instance, if we expect a curvilinear relationship between the independent variable and conflict behavior, as Mansfield (1994) argues regarding the relationship between system power concentration and the onset of war,
we must include a polynomial term to capture the effect. Alternatively, a “theory” might predict that one variable should be positive but another negative, as Snyder (2000) speculates regarding the relationship between democracy, democratization, and war. In this case, our test of Snyder’s argument consists of evaluating several variables simultaneously. This involves assessing the substantive effects of the block of two or more variables together, as well as the statistical significance of the individual variables. For instance, to test the dyadic variant of the democratic peace conjecture we estimate a model with the democracy level of a conflict initiator, the democracy level of the target, and the interaction of those levels. We examine the explanatory power of the variables by varying both the individual democracy scores as well as the interaction term.

What about the Possibility of Varying Effects across Different Levels of Conflict?

Commonly, theories built around rational expectations, selection effects, or signaling mechanisms contain careful deductive language, with assumptions and causal logic explicitly stated. Sometimes empirical tests of the causal mechanism do not lead to clear-cut results as we might expect to see in more simplistic conjectures. For example, when leaders strategically select their actions, we commonly see positive effects on the likelihood of conflict at one level of escalation and opposite or null effects at other levels. For example, Fearon (1994b) used this line of argument to explain why alliances appeared to have little effect on extended deterrence in Huth’s work (1988). Fearon argues that, when leaders considering an attack observed an alliance between a protégé and a defender but chose to challenge the status quo anyway, they must be particularly resolved, and so we would expect an insignificant relationship between the presence of the alliance and the success and failure of the extended deterrence attempt. In the context of theories built around signaling games, leaders facing severe domestic constraints or audience costs (e.g., democracies) may be quite willing to engage in conflict at low levels and simultaneously quite averse to participating in large-scale wars (cf. Senese 1997; Gartzke 1998).

To allow our analysis to be sensitive to these types of problems, we construct our dependent variable with multiple unordered categories that reflect various levels of conflict. As a result, with tests like the ones employed here, we can examine the effects of key variables representing theories across multiple levels of conflict.
Signorino (1999) points out another way in which imprecise or poorly specified theories and tests of them create implicit limits to knowledge. Signorino argues that most theories of international relations are actually theories of strategic interaction. That is, leaders make decisions about war while taking into account the likely reactions and responses of their opponents, either domestic political ones or international military ones. If this is the case, then simple decision-theoretic arguments about the presumed relationships between variables and international conflict are likely to be wrong and, even worse, may suggest including variables in a form that will lead to incorrect inferences about the variables’ effects. Two implications of this argument then follow. First, when testing these theories based on the logic of strategic choice, we must be very careful to include variables in a form that matches the theory’s internal logic (such as including game theoretic equilibria and not the expected utility components that factor into the equilibria). Second, our “theories” must be specified clearly enough to allow us to determine this form. Unfortunately, most so-called theories of international politics do not do this. Instead, they pose far looser or poorly specified relationships among vague concepts. In our analysis, we are unable to take each of the theories we include and redevelop their arguments in keeping with strict deductive logic. Rather, we take their hypothesized relationships and measures and include them “as is” for testing. If these “theories” suffer from flaws in their internal logic, or the operational variables do not accurately measure the causal influences, this may result in empirical findings that are inconsistent with the theories’ predictions.

How Meaningful Is It to Talk about “Relative Predictive Power”?

We have repeatedly discussed the relative power of different models, and comparing theoretical arguments based on their relative explanatory or predictive power. Some may wonder how this is possible. We discuss the details of our methodology for predicting the probability of conflict from our comparative model in the next chapter, but we provide a preview here to demonstrate that relative comparisons among competing explanations and conjectures are feasible and can be quite useful. As we have discussed, one or more variables represent each argument or conjecture in the models we estimate. We assess the effects of individual variables by focusing on the relative risk of conflict initiation and escalation, that is, the increase in the risk of conflict caused
when an independent variable changes by some known amount. For instance, if the probability of conflict in a dyad went from 1 percent to 3 percent when the democracy scores of the potential initiator went from highly democratic to highly autocratic, we would assess the increase in risk at “3x,” a threefold increase in the odds of conflict. For any variable in our model, we can estimate the increase in the relative risk of conflict in this manner. We can then make meaningful comparisons because the effects are in the same units (multiples of conflict risk); if changing one variable increases the risk of conflict three times and changing another increases the risk four times, then we will say that the second factor or argument has a relatively stronger effect.

The key to comparing the effects associated with the various propositions is to ensure that we are changing the values of the independent variables by a “comparable” amount (if we change one variable across its full range and another by a fraction of how it could change, then the comparison would be meaningless). This appears at first glance to be difficult, as there is no common metric for a wide set of independent variables measured using different scales. Democracy scores, for instance, are measured on a 20-point scale developed by the Polity project, while arms races are simply either present or absent, and the power concentration of the international system is measured on a scale potentially ranging from 0 to 1 (and in our data set between about 0.2 and 0.3). Nevertheless, even though the scales are different, we can still speak of “typical” changes in these variables based on the observed patterns found in the real world. For theories marked with dummy variables, this is relatively easy, as we can simply identify the no (0) versus yes (1) condition. With continuous variables, we use the mean as a typical value and use standard deviations to identify a range of real-world values that occur with some known probability. Given normally distributed variables, we would expect that, given a large enough sample, 95 percent of the cases in a set of data will fall within plus or minus two standard deviations of the mean. Most of our variables do not follow the normal distribution, so therefore the standard deviation does not capture precisely this number of cases for all variables. Nevertheless, it does give us some indication of the range of values we see in the real world. We can then use the standard deviation to suggest specific cut points in the data set where we will assess the estimated probability of war. In most cases we choose to change the values of the independent variables by two standard deviations, moving from one standard deviation below the variable’s mean to one standard deviation above the mean, a range that approximates a common set of occurrences. Note, however, that we are not using the once-common approach of present-
ing standardized regression coefficients. We focus on real-world changes in each variable to project outcome probabilities, but our unit of change is simply standard deviations for convenience as a measure of typical values. The fact that the variables use different scales is not a problem. By focusing on the range of data one standard deviation above or below the mean, our results remain general, so that we can look across theories and variables at the explanatory power in terms of relative risk. If one chooses the values for the simulation unwisely, of course, then the comparison may simply not be a very useful one. We agree with Clarke (2001) that we do not want to use this information to select the model with the greatest number of important variables and then present this set of independent variables as a better model of politics than another. We take exception, however, to Clarke’s (2001, 730) assertion that we cannot meaningfully compare the effects of several variables, given changes in independent variable values and outcome probabilities.

Comparisons of other variable shifts could be useful in other settings, particularly for policymakers choosing what actions to undertake in the context of a budget constraint. Many of the independent variables in the model are manipulable to greater or lesser degree. Some of the indicators do not vary in a manipulable sense at all—for example, polarity and geographic contiguity are simply descriptive facts, although in the long run polarity could perhaps be altered by conscious actions, such as European unification. Other variables are more directly manipulable by policymakers. For instance, leaders might be able to influence a country’s level of democracy (Przeworski et al. 2000) or take steps to expand trade with it. Changing a country’s level of democracy or trade might not be easy, or even certain, but an investment of a given size correlates with some probability of future regime change or increase in future trade flows. If we were to develop expected value estimates for each variable in our model, we could compare their relative effects based on the cost-effectiveness at reducing the risk of conflict.

Finally, we consider an analogy to our arguments concerning relative explanatory power, namely, the question of which is a more important influence on a child’s development, its environment or its heredity. If genetics were the determining factor in development, then we might logically resign ourselves to never improving our fate (future advances in gene therapy notwithstanding). Of course, in reality we expect that both environment and genetics play an interactive role in influencing the growth and achievement of children. Children who are brought up in loving homes, with adequate food and education, are likely to do better than those raised in less advantageous environments. Simultaneously, children with certain genetic traits are likely to
grow faster or suffer fewer ailments compared to those with a different genetic makeup.

When cast in such general terms as “environment” and “heredity,” the question of relative importance has no meaningful answer. Within the category of environment, however, specific and carefully measured environmental factors may have more or less powerful effects. If these factors, such as head start programs versus prenatal care, have comparable costs but differ greatly in their ultimate outcome, then assessing relative risks associated with the variance in manipulable factors such as a state’s level of democratization is a prudent strategy. Similarly, if genetic precursors of disease are associated with varying degrees of risk, and if we possessed a limited budget for genetic screening, we would want to focus on those factors associated with the highest risks where the hope of gene therapy was greatest. In the context of international relations, given that the actors are strategic, simply knowing that a given situation is potentially dangerous may lead decision makers to be more risk averse than they would be otherwise, thereby potentially averting conflicts we would otherwise expect to see. Absent this sort of comparative explanatory knowledge, advocating policy on efficiency grounds becomes next to impossible. In the next chapter, we turn to a fuller discussion of our selection of the theories, hypotheses, and conjectures we investigate, as well as our research methodology. In chapter 4, we will examine each of the arguments and the key variables that we derive from them. We then turn to our actual tests and to a comparison of the various arguments’ explanatory power.