Appendix A
The Rivalry Web Site

The data used for the various analyses in this book are available on the web site:

http://www.pol.uiuc.edu/faculty/diehl.html

The documentation for the web site and the data files contained therein can be found in the file "rivalry.readme." This file contains the description of the data found in the files "rivalry1.data, rivalry2.data . . . " All files are in ASCII format. We encourage those interested in part of the data to download the whole site (via the option provided on the web page), which is quite small, less than one megabyte.

The purpose of the web site is to provide the data used in this book for replication or further analysis. In most cases, the basic raw data are no longer the most current. Hence this site should *not* be used as a source for data on militarized disputes, regime type, and the like.

In some instances, particularly regarding descriptive statistics, the raw data provided need further processing before our results can be obtained. But in no case does this require extensive programming beyond calculating means, finding sums, and so forth.

The web site also assumes that one is in possession of this book. The file readme.txt provides only the basic data structure and variables with links to the tables of each chapter.

We welcome questions, comments, suggestions (and praise), which can be sent to either or both authors at

p-diehl@uiuc.edu

ggoertz@u.arizona.edu

Appendix B
An Index of Dispute Severity

In this appendix, we develop a new measure of dispute severity. There are several purposes in this exercise. More immediately for our needs, such a measure of dispute severity allows us to track hostility levels in rivalries. Disputes are the main signposts along the way that help define rivalries and their continuation. In chapter 3, we use the dispute severity measure to investigate whether some kinds of rivalries are more severe than others. An indicator of dispute severity is also important in developing an operational measure of the basic rivalry level (BRL), first explicated in chapter 9. Patterns in dispute severity over time help us determine whether that basic rivalry level is consistent over time, as predicted by the punctuated equilibrium model (see chapter 7), as well as whether conflict management or escalatory patterns of behavior are evident in a given rivalry.

Although the dispute severity indicator was constructed to serve our needs in this book, we also sought to construct an indicator with broader applications. Thus, we want a measure of dispute severity that has general validity, one that can be used in many theoretical contexts. In this sense, we follow J. David Singer's philosophy that good data sets can be used for multiple purposes. Most obviously, a good indicator of dispute severity is essential for studies that seek to predict conflict. The overwhelming majority of militarized confrontations do not escalate to war, and a more precise indicator of the relative severity of those disputes or crises would allow scholars to predict and explain the most and least serious of those conflicts. The implicit assumption in extant research is that these confrontations are largely indistinguishable. Furthermore, an interval measure would permit an analysis of the bargaining behavior of states and its impact on outcomes across disputes or crises. We could then assess whether conflict management strategies resulted in lower levels of conflict between states or whether the impacts were negligible. Currently we can only assess whether states are able to avoid war, and thus we have little sense whether they can manage conflict at lower levels.

In summary, this appendix has a simple goal: develop an indicator of dispute severity. Nevertheless, doing so forces us to reexamine and question basic

conceptualizations and practices. It turns out that the small act of indicator construction forces an examination of some basic principles and methodological practices. Our indicator rests on these basic ideas, to which we now turn.

Causes of War: Theory and Practice

The conceptualization of war as a dependent variable in conflict studies rests on theory, data, and methodological practice. These three are deeply intertwined and cannot be completely separated. Indeed, it is because they are so interrelated that our indicator construction project requires us to examine theory, data, and methodology. At a basic level, our proposal is to replace the war/no war distinction by a continuous measure of dispute severity that includes war. This suggests that the war/no war dichotomization misses important aspects or misrepresents key dimensions of the phenomenon of international militarized conflict.

When the COW project decided to base its dichotomous definition of war on the number of military fatalities, the tension inherent in the contrast between fatalities, which are continuous, and war, which is categorical, existed. Of course, one very early—and easy—critique was that there are cases of war that fall just below the one thousand-death cutoff, or some nonwar incidents that land above (Duvall 1976). For example, the Battle of Savarino Bay is not a war, while the Soviet-Japanese skirmishes in the late 1930s were coded as wars. Indeed, this is partly the basis for a reformulation of the war data set by Siverson and Tennefoss (1982), who adopt a lower threshold of fatalities. Even the critics, however, classified war in the same dichotomous terms. The disagreement was over the fatality threshold level or its precision and not over the dichotomous conceptualization of war.

We have no objection to referring to very serious militarized disputes as wars, but the question is the degree to which we want to formulate scientific hypotheses in those terms. To choose an analogy from physics, in common parlance we talk about hot and cold water. Yet, as scientists, we want to use temperature as our dependent variable. This has important implications for the framing of hypotheses and the interpretation of statistical methods. For example, the democratic peace is framed in terms such as "democracies do not fight wars," not in terms of a hypothesis about how severe disputes between democracies are. Rummel (1995) indicates the democratic peace is most evident with respect to violence severity, a claim that cannot be properly tested with a simple war/no war analysis. The dichotomous conceptualization then produces strained discussions of whether Finland–United Kingdom in World War II counts as a "war" between democracies (Russett 1993; Ray 1995).

If we examine methodological practice, a related set of issues arises. For example, given the dichotomous dependent variable, most studies use event history statistical methods such as probit and logit. These statistical methods

provide a continuous predicted value between 0 and 1. How is this usually interpreted? The answer is virtually always in terms of the probability of war; for example, a predicted value of .70 means that the predicted probability of war is 70 percent. With our continuous indicator and concept, as well as the adoption of different statistical methods, the predicted value is the level of dispute severity. As the logic of the democratic peace applies to dispute severity as well, such a dispute severity dependent variable would allow us to determine where the threshold of severity for the democratic peace really lies. It may be that it falls significantly below all wars, a finding that is not easily discernible using the war/no war classification.

Another tension lies in the conceptualization of war as a potentially multilateral event and the practice of statistical analysis of *dyads*. For example, Vasquez (1996) has argued that there are two paths leading to war, one through dyadic rivalry and the other through multilateral contagion. As we shall see below, the practice of "dyadization" of multilateral wars proves quite problematic in most works using MID data. It turns out that many dyadic relationships within a multilateral war do not deserve the label *war,* and in fact are not coded as such in the MID data set itself. For example, the U.S. participation in the Opium War was not a war from the American point of view, but certainly was from the Chinese standpoint; correspondingly China is assigned a war code, while the United States gets the use-of-force coding.

Another hidden implication of the dichotomous war/no war coding arises from the use of event history methods. These models are all nonlinear. Thus, for example, when analyzing the impact of a given variable, one normally holds the other variables either at the extremes or at their means. If one were to have an interval level dispute dependent variable, almost certainly researchers would use linear regression. Rarely are conflict theories explicit on the linear versus nonlinear question (Gelpi 1997). Most of the time this occurs as a side effect of the methodology: linear for regression and nonlinear for event history techniques. Yet some theories are implicitly nonlinear. For example, necessary condition hypotheses (which include the democratic peace and the power transition hypotheses) imply nonlinearity, and probit and logit models do not accurately test these necessary condition theories (Braumoeller and Goertz 1997). It is certainly the case that one can test linear versus nonlinear models more easily with a continuous dependent variable. Here the dichotomous dependent variable induces through the event history methodology a nonlinear model. In principle, it should work the other way around: the theory should determine the functional form of the statistical model, not the measurement of the dependent variable. It is easily forgotten, particularly in the statistical testing literature, that "theory" means more than a particular collection of variables, that it also should include functional form.

By moving from an exclusive focus on war to dispute severity, we expand dramatically the range of phenomena we can examine, which itself is a spur to

theoretical development. One of the current interests in international conflict studies is conflict management and resolution. If we want to investigate and evaluate peacekeeping and mediation success, for example, we need a much finer-grained measure of dispute severity. One of our interests (see the second part of this book) is the evolution of rivalries (see also Diehl 1998). Yet it is all but impossible to track changes in rivalry dynamics without a more nuanced look at dispute severity; by focusing only on the war/no war distinction, scholars will miss trends toward increasing or decreasing (e.g., conflict management) interactions.

The core of our severity measure goes back to the original idea of the COW project: the severity of a militarized conflict is very closely related to the number of fatalities. Because we are not concerned with "war" per se, we need not worry about a strict demarcation between wars and nonwars. Of course, many problems remain. We now turn to the theoretical and practical aspects of developing a measure of the concept of dispute severity.

Previous Efforts at Measuring Conflict Severity

There have not been extensive efforts to develop interval measures of conflict severity in the international conflict literature, largely because the concern has been with understanding the conditions for war and not with more subtle variations among conflict events. In making a distinction between severities of conflict, we begin with the war/no war distinction that dominates the literature. The Correlates of War Project (Small and Singer 1982) developed the most widely used measure, although there are also other classic formulations (Richardson 1960; Wright 1965). Generally these efforts have used the number of battle-related fatalities to distinguish wars from other conflict; for Small and Singer the threshold for war is set at one thousand deaths.

The war/no war distinction has some significant shortcomings as a basis for assessing conflict severity (see Duvall, 1976). Most obviously, the simple dichotomous classification ignores enormous variations within each of the two categories. For example, there is no distinction made between crises that involved a significant use of force and fatalities on the one hand and seizures of fishing boats on the other. Furthermore, relatively minor wars such as the Football War between Honduras and El Salvador are lumped together with major conflagrations such as World Wars I and II. In addition, the Correlates of War Project list of militarized disputes, which includes disputes that went to war as well as those that did not, has some cases in which some participants are coded as having gone to war while other participants did not reach the war level. How does one categorize a conflict when some participants get a war coding and others do not? At the aggregate level, a conflict can be categorized as a war if only one disputant reaches the appropriate level of hostility or fatalities to be coded as a war. Furthermore, a state may have exactly the same number of fatalities in two different conflicts and one can be coded as a war and the other not. This makes the use of even the simple dichotomous classification problematic.

Similar to the distinction between war and its absence are attempts to distinguish whether a conflict is a crisis or not (Wilkenfeld, Brecher, and Moser 1988). An international crisis occurs when there is "(1) distortion in the type and an increase in the intensity of disruptive interactions between two or more adversaries, with an accompanying high probability of military hostilities, or, during a war an adverse change in military balance, and (2) a challenge to the existing structure of an international system . . . posed by the higher than normal conflict interactions" (Wilkenfeld, Brecher, and Moser 1988, 3). Unlike the criteria identifying war events, the ones used to distinguish between crises and other events include some relatively intangible variables. Not only does this make it difficult to again draw a dichotomous distinction, but it is not very helpful in providing any basis for constructing an operational, interval-level measure.

The International Crisis Behavior (ICB) data set makes distinctions between the severities of different crises. The ICB data set includes data concerning several dimensions of crisis severity. Severity is assessed by (1) the number of actors involved, (2) the level of involvement by great powers, (3) the geostrategic salience of the conflict, (4) the degree of attribute difference (military economic, political, cultural) between the participants, (5) the number of crisis issues and the presence of military security issues, and (6) the extent of violence in the crisis. The data set includes an index of crisis severity that is a weighted summation of the six dimensions of severity. The weights assigned each indicator are based on the number of postulated linkages between one dimension and all the others. The overall severity scores are then transformed to a 1–10 integer scale. Although this index provides an aggregate measure of crisis severity, it is not clear that it is truly interval level and it still suffers from the shortcoming that it is constructed largely independent of the actual behavior of the crisis actors. Only the violence dimension refers to any behavioral characteristics of the crisis in defining its severity. The remaining dimensions characterize the attributes of the participants or the substance and location of the crisis. The ICB method of measuring severity appears to confound the *causes* of severity from the severity dependent variable itself. For us, the first five factors would typically be part of the explanation of crisis severity, not crisis severity itself.

Other crisis data sets (Leng and Singer 1988) do not contain a specific index of severity. Rather, there is a description of individual events or actions within the crisis (e.g., verbal threat and surrender) that may provide the basis for comparative scaling in terms of severity or hostility. Indeed, such data have been used to explore the various coercive or reciprocal bargaining strategies within crises (Leng 1993). Yet it is not clear how one would aggregate these events (more than 25 thousand events across 40 crises in the Behavioral Correlates of War—BCOW—data set) to formulate an overall severity measure.

The Correlates of War MID data set includes measures of the highest level of hostility (LOH) reached during the course of the dispute by each of the participants. The scale is an ordinal one, ranging from 1 to 5, where 1 is no military response (relevant only for the target state), 2 is threat of military force, 3 is display of military force, 4 is use of military force, and 5 is full-scale war (Gochman and Maoz 1984; Jones, Bremer, and Singer 1996). Goertz and Diehl (1998) used a multiplicative scheme using the LOH scores of each rival to indicate conflict severity. Crescenzi and Enterline (1998) go one step further in constructing a "rivalry" severity score for each dyad; their technique uses the LOH score for disputes along with a time decay function. Yet their measure may be suitable when rivalries or rivalry years are the units of analysis, but the measure cannot be easily disaggregated to the dispute level although it uses a dispute severity indicator (the level-of-hostility variable) as one input. There are several limitations to the simple LOH scale. By definition "militarized dispute" limits the consideration to events that involve at least the possibility of military force and therefore does not distinguish between hostile and coercive actions that do not involve military force. Although this scheme is a useful refinement of the basic dispute/war classification, there are still problems with using it for severity. First, the five-point scale is only ordinal, incapable of distinguishing the magnitude of difference across the categories. Second, the scale does not recognize the huge differences (in costs, duration, etc.) between simple disputes and wars; indeed, war is considered only the most severe form of dispute. Third, there is little consideration of the relative symmetry of state behavior. Traditionally, conflict studies define dispute severity as the highest level of hostility achieved by *any one state in the dispute*. This codes as equivalent cases where both sides use force (i.e., 4-4 dispute) and those where one side made no militarized response (4-1 dispute). Finally, the scale does not distinguish between different war severities, therefore offering little improvement over the simple war/no war distinction.

The most recent edition of the Correlates of War dispute data set (Jones, Bremer, and Singer 1996) contains a 22-point scale for militarized action in a dispute, given in table B.1 (for an application using this scale as well as fatality levels, see Senese 1996). Although there are some helpful distinctions between actions that fall within the same category on the original five-point scale, it is not clear that the scale is even of ordinal character, much less an interval measure. Is a show of troops inherently less severe than a show of ships or planes? Are threats to blockade (technically an act of war under classical international law) different from the threat to occupy territory or go to war? One might clearly say that a nuclear alert was more serious than a simple military alert, but is a nuclear alert less severe than a mobilization or other action? These rankings are not obvious, and certainly context-specific factors may lead to a dramatic reordering of actions on a hierarchy of severity. Most critically, according to Stuart Bremer (personal communication), COW only recorded the

TABLE B.1: MID Hostility Scale

5-Level Scale	22-Level Scale
1 = No militarized action	1 = No militarized action
2 = Threat to use force	2 = Threat to use force
	3 = Threat to blockade
	4 = Threat to occupy territory
	5 = Threat to declare war
	6 = Threat to use nuclear weapons
3 = Display of force	7 = Show of troops
	8 = Show of ships
	9 = Show of planes
	10 = Alert
	11 = Nuclear alert
	12 = Mobilization
	13 = Fortify border
	14 = Border violation
4 = Use of force	15 = Blockade
	16 = Occupation of territory
	17 = Seizure
	18 = Clash
	19 = Other use of force
	20 = Declaration of war
	21 = Use of CBR weapons
5 = War	22 = Interstate war

first act within the narrower five-point hostility scale, rather than necessarily the highest act (1–22 scale within the categories); for example, if a blockade (15) by one state occurred before an act of seizure (17), only the blockade is recorded in the MID set, as both fall under level of hostility 4. This compromises the use of the 22 point scale as an accurate representation of dispute severity.

Maoz (1982) constructed an interval measure of dispute severity from an early version of the COW dispute data set. Using a 14-category scale of actions (a middle ground between later five- and 22-point categorizations), the temporal order in which various "incidents" occurred that make up a dispute, and data on these individual incidents, he constructed a dispute-level measure of severity. In some ways, the Crescenzi and Enterline (1998) measure does for rivalries what Maoz (1982) did for disputes. In each case, individual incidents or disputes and their occurrence over time are used to construct a severity indicator.

Wang (1995) developed an interval scale of U.S. responses to foreign policy crises, using insights from the both the COW and ICB data collections. He created 11 categories of responses scaled from 0 to 1 with compliance and external violent military responses representing the end points. Theoretically, one could construct a similar scale for crisis initiators or other participants, but it is not clear that all such categories would be applicable (compliance, for example, is an inappropriate category for an initiating state), and there is still a question of how to aggregate the scores of the different crisis actors.

Beyond the Correlates of War and International Crisis Behavior Projects, others have constructed hostility scales. COPDAB (Azar 1982) and WEIS (McClelland 1976) are events data collections that classify specific actions of states on 16-point ordinal and 60-point nominal scales respectively. One benefit of these schemes is that cooperative actions as well as conflictual ones are recognized. Because they are events data compilations, there are considerably more actions than are present in the MID data set. Thus, a measure of dispute severity would require an aggregation of some or all relevant events within the appropriate time frames; in some cases, a useful time-series must be constructed from the raw data (see Goldstein 1992 for a technique to do this with WEIS data). There are also several difficulties with using events data. Most obviously, events data are available only since 1945 (and then not necessarily for all countries), making longitudinal analyses of some dyads or rivalries impossible and precluding analyses of conflicts before World War II. Second, good reasons exist to believe that COPDAB and WEIS compilations may not be reliable or valid (Howell 1983). Third, the aggregation of many different events may obscure the salience of some key ones, most likely those involving military force, and the aggregated measure may underestimate the severity of the confrontation.

Finally, some scholars have argued that disputes over certain issues are inherently more dangerous or severe than other disputes, with territory being the issue most often cited (Vasquez 1993). Thus, it might be argued that some attention be given to issue in noting the severity of a conflict. We reject this implication. The issue of a given dispute is perhaps a valid predictor of conflict or escalation, but it should not be used as an indicator of the severity of that conflict. We must not confuse causal factors associated with conflict severity, with measures of the magnitude of severity itself.

The Underlying Dispute Severity Concept

In developing an index of dispute severity, we must pay attention to the underlying concept that we are trying to measure. It seems clear that if we define *severity* as "degree of military force used," then threats are less severe than displays of military force, which in turn are less severe than uses of force. Nevertheless, *severe* has other connotations. For example, most analysts consider the Cuban missile crisis as the most severe crisis between the United States and the USSR,

yet this only gets coded as 4 (15) for the United States and 4 (19) for the Soviet Union in the MID data. This crisis may have been more severe than many uses of force, which receive a higher coding in the COW level-of-hostility scheme.

It is clear that severity does not always equal "degree of military force used," at least in the intuitive sense. Here, severity appears to indicate "risk of war." This is the implicit notion of severity that arises using event history methodologies (see above). We think that trying to explain the probability of war or evaluating dispute severity in terms of the risk of war constitutes a valuable, but different, research enterprise. One practical problem with using dispute severity defined as risk of war is that there is no independent evaluation or data for this. The estimates produced by event history techniques obviously depend on the model, data, and indicators used in the given study; one can hardly use these as relatively "theory neutral" measures across a wide range of topics. Part of our goal is an indicator, similar to the COW definition of war, that is usable across a range of problems and theoretical perspectives.

It is important to understand that all scales—including those in the natural sciences—include theoretical and empirical considerations. The COW level-of-hostility scale is no exception. On the one hand, obviously theoretical notions about what constitutes increasing levels of military force are taken into consideration, but that is only part of the story behind the scale. Richard Stoll reports (personal communication) that when coding disputes, researchers at the COW project noted that the scale ordering was not only a level of force 1, but also represented a common temporal order of "escalation" (this word itself implies some sort of scale) of a crisis or war. Maoz (1982) quite explicitly used an analogous temporal ordering principle to construct his interval scale.

Thus, before moving to the details of our proposed measure, we need to ask how we conceptualize severity (we will henceforth use the term *dispute* to refer to both nonwar and war disputes, unless the context indicates otherwise). There are a number of general principles that guide our development of a measure of dispute severity.

First, we look for a unidimensional measure of dispute severity. For example, events data use a cooperation-to-conflict scale, implicitly signifying *one* dimension.

Second, we believe that severity increases with the level of military threat or force.

Third, we see dispute severity ideally as a scale that increases with the level of *actual* military force. It should be explicitly stated that we are looking for a *behavioral* measure. It is quite possible that the level of hostility at the psychological level does not correspond to the level expressed in action. Here it is important to keep in mind that we want to use the dispute severity measure most often as a dependent variable. Hostility as a psychological variable is more often used as an independent variable (i.e., as an explanation of behavior). Thus, we can contrast the actual severity level against the risk-of-war perspective on

dispute severity. In well-defined situations, we can talk about "objective risk," such as in classic probability, gambling, and natural science applications, but it appears almost impossible to develop something analogous for international war.

Fourth, another important distinction is that we are trying to determine the severity level for the dispute between two countries: the unit of analysis is the dyad. The dyad is the preferred unit of analysis for dispute severity given its predominance in studies of international conflict and its flexibility in application; with respect to the latter, one can aggregate dyad scores for a multilateral dispute score, but the reverse is not true. Obviously, this dispute-level variable will be constructed with data about what each participant itself has done. Even at the simplest level, however, one has to make a decision how to combine the data from the two parties into a dispute-level measure: should one add, multiply, or take the maximum of the two disputants' scores? Thus, we have at least three different options, each of which produces quite different results. We discuss these options in the next section.

Related to the fourth concern is a preference for a measure that captures the symmetry of the conflict level achieved by both rivals, the fifth general measurement principle. Frequently, the conflict level of militarized disputes has been measured by reference to the most severe acts of military force committed by one state, ignoring that the other party may exhibit a much less hostile reaction, or indeed no response at all. Some militarized disputes involve no military reaction by the target state after the initiating state threatens or uses military force (Hensel and Diehl 1994). Thus, the severity level should be greater when military actions are met with reciprocity as opposed to less hostile reactions. This is consistent with the dyad being the unit of analysis.

Sixth, we prefer an interval measure rather than the nominal or ordinal ones that generally characterize past efforts. An interval measure permits more precise conclusions to be drawn about severity and opens up a much broader range of statistical techniques and theoretical models to use in understanding dispute severity and its correlates.

Seventh and pragmatically, we are confined to constructing an indicator for which we have the necessary data. Variables that already exist in the COW militarized interstate dispute (MID) data set (Jones, Bremer, and Singer 1996) as well as the COW war data set will form the bases of our efforts.

In short, we are looking for (1) a unidimensional measure (2) that increases (monotonically, but not necessarily linearly) the level of force, (3) that uses the actual military force used or its effects, (4) that describes the dyadic dispute as a whole, (5) that reflects the symmetry of dispute participant behavior, (6) that is interval level, and finally and pragmatically, (7) is one for which we have data. Notice that unlike the Maoz (1982) scale or the simple COW level-of-hostility scale, we do not assume that this scale represents a temporal order of events.

A Measure of Dispute Severity

If we examine the basic idea behind the Small-Singer (1982) war data set, we find that the number of battlefield fatalities effectively indicates the severity of militarized conflict. They defined war as incidents that result in one thousand or more fatalities. We extend this to the principle that the severity of a war or dispute is a function of the number of fatalities, and therefore dispute severity will be, partly, a function of the number of deaths that occur in each conflict incident. This is suitable for disputes that actually have fatalities, but the large majority of militarized disputes involve no deaths. Clearly among this larger group, some disputes are more severe than others are. Hence, we need another procedure for measuring severity in nonfatality cases. The MID data set does provide information on severity levels for nonwar cases; this is contained in the level-of-hostility variable, an ordinal variable ranging from 1, no response, to 5, war, described above.

Our first operational principle in constructing a severity measure from all disputes and wars is therefore this: if fatality levels are greater than zero, then the severity level is a function of those battle deaths. The second principle is that if the fatality level is zero, then the severity level is a function of the level-of-hostility variable for the dispute. A third principle is that disputes with fatalities are more severe than disputes without deaths. We recognize that some conflicts are very serious and nearly lead to war even though there is no loss of life and that a few conflicts inadvertently result in bloodshed and death but do not pose the same risk of war. Nevertheless, the loss of life has strong symbolic, substantive, and domestic political implications and necessarily conditions the perceptions and accompanying responses of decision makers. For these reasons, we regard fatalities as a key component of our indicator of dispute severity.

Having specified our general approach, several problems remain to be solved. We must first construct interval level measures for each part of the scale, the zero- and nonzero-fatality segments. Then we must devise a mechanism to splice them together to form one overall measure that is comparable across different disputes.

We begin with the zero-fatality portion of the scale. The first question is how to combine the data from the two level-of-hostility variables (one for each disputant) to form a dyad-level value. Three simple options exist: (1) addition, (2) multiplication, and (3) the maximum of the two. Currently, almost all users of the MID data set take the maximum option, that is, the severity of the dispute is indicated by the highest level of hostility achieved by any party to the dispute. We think option 2, multiplication, makes the most sense. This alternative appears to deal with asymmetrical cases (those in which the rivals do not reach the same level of hostility) in the best fashion. To take the extreme example, one rival can reach a level of hostility of 5 (war) while its opponent

only reaches the second level (threat).[1] To indicate severity by taking the maximum of the two scores means this dispute is treated the same as any other war, but we think that it is less severe because the other side exhibited a relatively weak reaction. This is especially important in large, multilateral disputes. We believe that the maximum option does not fulfill the reciprocity criterion noted above. We are especially attracted to the multiplication option because the differences in scores become larger the higher the level of hostility achieved by *both* sides. Thus, our first step is to create a new scale of dispute level severity for zero-fatalities cases by taking the product of the level-of-hostility scores for each rival.[2] This then ranges from 2 (2×1) to 16 (4×4).[3]

Figure B.1 shows the distribution of cases (remember that this is for zero-fatality disputes only). The values of 4, 9, and 16 are the symmetrical nonfatality disputes, and all other values are nonsymmetrical disputes. Clearly the symmetrical cases dominate, but there are a significant number of nonsymmetrical ones.

A second step is to convert this ordinal scale into an interval one. We adopt an inductive approach and suggest that the interval scale be a function of how frequently the different levels (2–16) occur in practice. The scale thus depends on the empirical facts, but this is true of scales such as temperature, which is defined based on the behavior of substances. We propose that the relative frequencies of each level be indicated by the cumulative distribution function in this particular case, which we give in figure B.1. We rescaled the severity variable using the cumulative distribution at any given point to define the new value. For example, the original ordinal score of 4 (2×2) now has a value of 58. Cases of reciprocated use-of-force (i.e., 4×4 = 16) get a value of 100.[4]

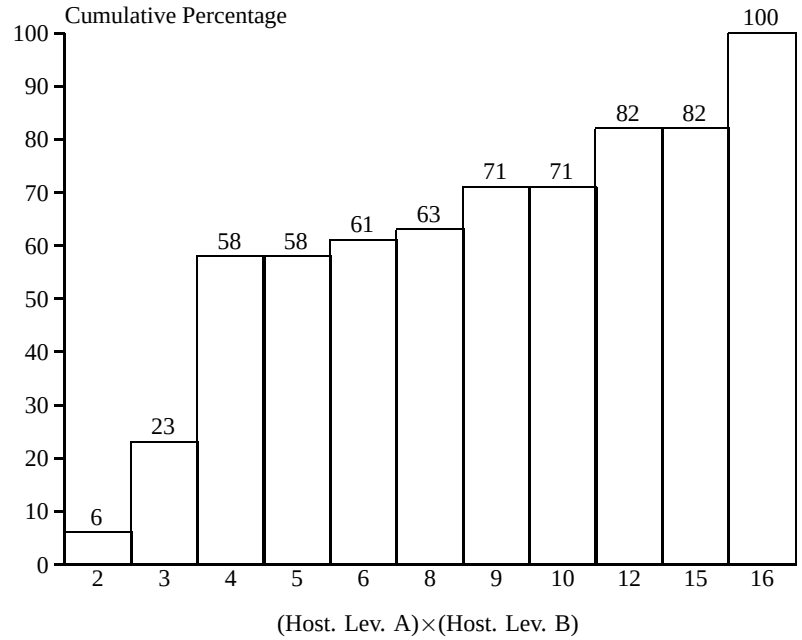
[1] There are actually cases of this sort in the data set. Indeed, they are likely to be more prevalent when there is a multilateral dispute and some participants do not partake in actual warfare while others do.

[2] We must deal with the problem of missing data. Only a very small number of hostility scores are missing for the initiating side in the dispute, but just under half are missing for the target. We suspect that much of the missing hostility data consist of cases in which the disputant made no military response. In contrast with the old dispute data, the revised MID data set has very few cases where the level of hostility equals 1 (see Hensel and Diehl 1994). That almost all the missing data involve side B would tend to confirm that. Nonetheless, there are virtually no missing data for the dispute level variable, the variable that most studies use. Implicitly this suggests that the missing data for the dispute target are considered lower than that of initiator and are probably cases of no military response. And indeed Stuart Bremer (personal communication) confirms that these are indeed no military response cases. Accordingly, we have set the missing cases to the level of 1—no response.

[3] The lowest number possible is 2, as this reflects the initiation of a militarized dispute with a threat of military force and a nonmilitarized response by the target. A score of 1 for both sides is technically impossible, as there would be no dispute to begin with: militarized disputes require at least one state to threaten, display, or use military force. Yet missing data for the initiator side leads to a small number of cases (N = 2) have a score of 1-1 for both sides.

[4] One will notice some values of 15, which are war (5)×display (3), and these are coded as no fatality cases; see below.

FIGURE B.1: Cumulative Distribution of Dispute Severity, Nonfatality Cases



If we examine the results in figure B.1, we think that they have a fair amount of face validity. We note that there are relatively few cases of LOH 2 with no response (2-1 dispute), which we regard as not very severe disputes. The move from nonreciprocated verbal (2-1 disputes) threats to nonreciprocated physical ones (display of force, 3-1 disputes) is indicated by the significant jump from 6 to 23. There is another significant jump then to level 4. This level includes reciprocated threats (2-2 disputes), which we consider much more severe than the nonreciprocated ones below it. Hence, low-level, nonreciprocated disputes on our scale have low values; it when we have a reciprocal serious threat that severity increases to over 50. LOH 4 also includes uses of force that have no response (4-1 disputes). We consider that it is much better to code these at this lower level than the standard approach that treats them as equivalent to level of hostility 4-4 disputes.

If we move up to level 9—reciprocated displays of force—cases there receive a value of 71. This is a more moderate increase—about 12—from the reciprocated serious threat level of 58, which to us indicates that the movement from a serious verbal threat to a serious physical threat is smaller than the jump to reciprocated verbal threats at level 50. The scale then moves up gradually to where a display of force is matched with a use of force (level 82). Finally,

there is a significant jump representing a reciprocal use of force (to 100), again emphasizing the importance of symmetry in severity.

In summary, we believe that within the limits of our data this part of the scale fits well with our theoretical considerations about severity as well as our concerns about the impact of asymmetry. It provides the kind of reasonable estimation and face validity that is similar to the use of one-hundred-point "feeling thermometers" in studies of American public opinion.

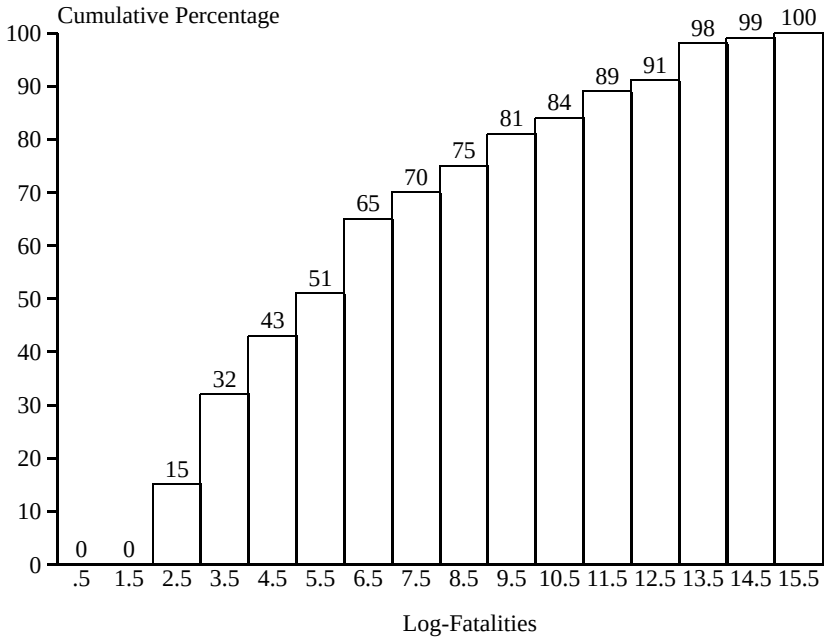
We now turn our attention to greater than zero fatality cases. Note that not all these cases are wars (and therefore some have dispute level of hostility codes of 2-16, but these are not include in figure B.1, which contains only the results for zero-fatality disputes). Both the MID data set and the COW war data set contain information about fatalities. For war cases, we used the more precise fatality numbers from the war data set. This involved matching wars from that data set to disputes in the MID data set. The MID data set categorizes them as follows: (1) 0, (2) 1–25, (3) 26–100, (4) 101–250, (5) 251–500, and (6) 501–999. If there were no precise fatality data from the war data set (i.e., cases of militarized disputes with fatalities less than one thousand, but greater than zero), we used the midpoint of these MID ranges as the fatality estimate for that disputant. The fatality level for the dyadic dispute is then the sum of the fatality levels of the two disputants.

The fatality level data in the war and dispute data sets are the total number of fatalities against all disputants. When we divide multilateral wars into dyadic cases, we need to consider whether we should take the aggregate dispute/war total for all the dyadic disputes.[5] In general we do not find this a problem, as most disputes and wars involve only two states, but it does become an issue in major multilateral wars. There, fatality levels can range from a few hundred to a few million in the same war. In fact, only the two world wars pose problems. Our solution consists of taking the minimum fatality level of the two states and multiplying that by two. This means that we consider the severity of the dispute confrontation between the two states to be best reflected by the fatalities on the smaller side. For example, we can take the extreme case of Brazil-Germany in World War II, in which Germany lost 3.5 million soldiers and Brazil only one thousand. We think that twice the Brazil fatality reflects a reasonable estimate (given our data constraints) of the severity of the confrontation between Germany and Brazil. Certainly this is much less severe than WW II conflicts between Germany and its other opponents in that conflict.[6]

[5]The transformation of multilateral disputes into dyads represents all possible dyad combinations from those states on the initiator side against those on the target side. Nevertheless, we eliminated some pairs in World War I, World War II, and the Gulf War. This was only done, however, when there was no temporal overlap between the war or dispute participants. That is, one state exited the dispute or war before the other state joined the conflict.

[6]For the fatality data we also need to take into account the problem of missing data. In the war data set, there were missing data for some participants in the Gulf War, namely Canada, Italy, Morocco, Syria, Bahrain, Qatar, and Oman and in some disputes (about 12 percent). As with the

FIGURE B.2: Cumulative Distribution of Dispute Severity, Fatality Cases



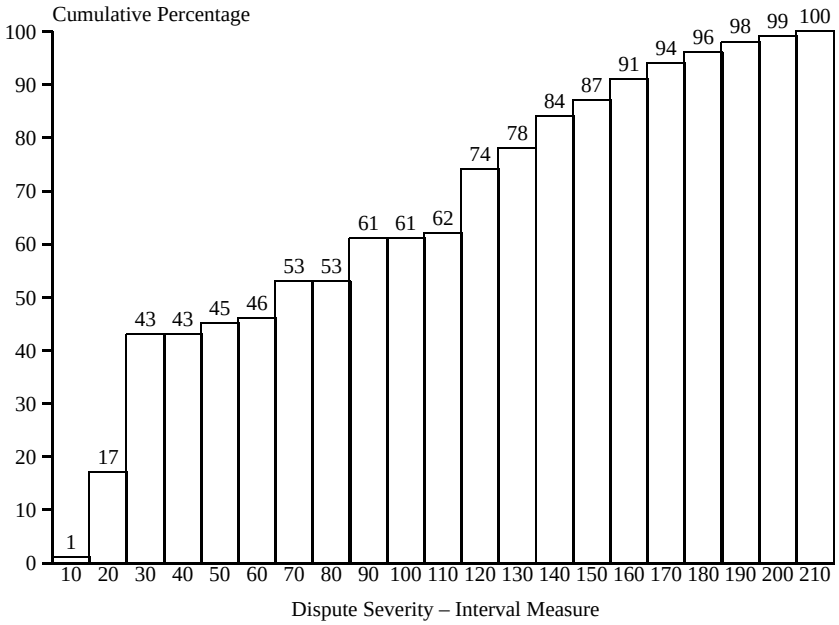
Because the absolute number of fatalities has an extremely long tail (few wars with very large fatality levels), we take the natural log of fatality levels to mitigate the effects of those outlying values. If we look at the distribution of these log values, it has a maximum of about 15.

Our proposal to scale the nonzero-fatality cases parallels what we did for the zero-fatality cases; the interval measure is the cumulative distribution of logged fatality values, which we show in figure B.2. Again to evaluate face validity we note that the scale increases fairly rapidly with small log-fatality levels. We think this reflects that the first fatalities substantially increase the severity more than do later ones. The first one hundred fatalities are much more important than those from 1,000 to 1,100. We think that the long tail thus represents quite well the slowly increasing severity of a war once a significant fatality is reached.

The next step is to splice the nonfatality and fatality case parts of the scale together. Based on our third general principle, we put the fatality case scores right after the nonfatality ones, and therefore the overall severity measure ranges from 0 to 200, 100 coming from each part. This means, in practice, a long tail for the fatality cases, which constitute half of the scale, but only

level-of-hostility missing data, we set these values to zero, and thus the cases are scaled with the zero-fatality disputes.

FIGURE B.3: Cumulative Distribution of Interval Severity Scores



about one-fourth of the cases. To make the calculation easier and to smooth things out, we fit a polynomial regression equation to these data. This permits one to calculate the interval measure based on very simple procedures. The fit of the regression line is quite good with an R^2 of over .99.[7] The equation is

dispute severity = 7.779×level – .034×level-squared,

where "level" equals either the product of the level-of-hostility scores (2–16) in the nonfatality dispute dyads or the log of the sum of the fatalities plus 16. Level-squared is the square of "level." The net effect of using the formula is that a few values will fall just beyond the 200 limit, but these scores can either be scaled back to the upper limit or accepted as is; in either case, empirical analyses should not be significantly affected. In summary, the procedure for calculating the measure is relatively simple. If there are zero fatalities, then multiply the level-of-hostility variable for each rival and apply the equation. For cases with battle deaths, the fatality levels for both states are summed, then logged, then added to 16, and then the equation is applied. Figure B.3 shows the distribution of the final scores.

[7] We checked various other specifications with higher-order polynomials, which never produced a significantly better fit.

Overall, we sought a scale that permits us to develop and test hypotheses about the gamut of dispute severity. The scale has a clear transition point at 100, which separates fatality from nonfatality disputes. Wars in the Small and Singer sense of one thousand fatalities begin around 160 and continue to a little over 200. Because much of our interest in developing this scale focuses on nonwar disputes, we have range from zero to approximately 150–160, which we find adequate for trying to differentiate between various dispute outcomes. We find that jumps in the scale occur where we would expect them to, and that asymmetric disputes are also properly treated. The rapid increase in severity with initial fatalities followed by a long tail also fits our notions about how severity increases with number of battlefield deaths.

Our justification of the face validity of our measure follows that of the Small and Singer data set. In the final analysis the justification for the choice of one thousand fatalities was that such a cutoff produced a list that they (and others) believed were wars and excluded those that most scholars would hesitate to identify as a war. We argue similarly that our scale reflects reasonable expectations about how dispute and war severity increase with the level of force, dispute asymmetry, and number of fatalities. If our expectations and arguments about how these factors relate to dispute and war severity are valid, then we feel that our measure also can serve a useful function since the scale fits well with those general arguments. Of course, if our general theoretical arguments about level of force, dispute asymmetry, and number of fatalities and their relation to severity are flawed, then of course so is our scale.

Conclusion

It would be easy to take an interval-level measure and merely change one's statistical techniques from logit to regression without giving the issue any further thought. But using our interval-level measure (or any other one, for that matter) consists of more than doing the same old thing with a new dependent variable. Behind the war/dispute distinction lies the basic orientation that one's goal is to explain war. Within this perspective, disputes are merely the control group. The shift to a continuous variable implies that the theoretical orientation shifts as well. Now, one is trying explain *both* disputes and wars in the sense that the distinction between the two has disappeared into a general measure of conflict severity. To take the analogy from temperature scales, the standard goal is to explain why something is hot, and this is replaced by the purpose of explaining its temperature.

It would be unfortunate to reduce our argument to one about information loss resulting from the war–no war dichotomization. There are important theoretical issues at stake. To explain war in traditional studies, strictly speaking, means explaining why some disputes reach the one-thousand-fatality threshold. The severity of the war once this threshold attained is treated as irrelevant in most studies. If we take our measure instead, then one will be explaining why

the war or other conflict became as severe as it did. We propose that a dispute severity measure permits us to analyze whole new sets of theoretical questions and at the same permits us to think anew about traditional procedures. War remains an important concern of conflict studies, but much conflict does not attain that level. We see the existence of a dispute severity indicator as an incitement to develop theories and to examine lower-level conflict.