Prediction, Proxies, and Power*

Robert J. Carroll† Brenton Kenkel‡

March 7, 2017

Abstract

Many enduring questions in international relations theory focus on power relations, so it is important that scholars have a good measure of relative power. The standard measure of relative military power, the capability ratio, is barely better than random guessing at predicting militarized dispute outcomes. We use machine learning to build a superior proxy, the Dispute Outcome Expectations score, from the same underlying data. Our measure is an order of magnitude better than the capability ratio at predicting dispute outcomes. We replicate Reed et al. (2008) and find, contrary to the original conclusions, that the probability of conflict is always highest when the state with the least benefits has a preponderance of power. In replications of 18 other dyadic analyses that use power as a control, we find that replacing the standard measure with DOE scores usually improves both in-sample and out-of-sample goodness of fit.

Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: http://dx.doi.org/10.7910/DVN/FPYKTP.

Word count: 9,244, plus 1,338 in Appendix.

*We thank Scott Bennett, Brett Benson, Bill Berry, Inken von Borzyskowski, Kevin Clarke, Josh Clinton, Mark Fey, James Honaker, Zach Jones, Karen Jusko, Holger Kern, Ashley Leeds, David Lewis, Adeline Lo, Matt Pietryka, Marc Ratkovic, Jim Ray, Mark Souva, and Hye Young You for helpful discussions and advice. We also appreciate helpful suggestions from the editor and three anonymous reviewers. Bryan Rooney provided excellent research assistance. We thank the authors listed in Table 6 for making their replication data publicly available.

†Florida State University. Email: rjcarroll@fsu.edu

‡Vanderbilt University. Email: brenton.kenkel@vanderbilt.edu

For all its progress—more nuanced arguments, more useful theories, bigger data and more systematic ways to analyze them—international relations remains, in many ways, a study of power. This is best reflected in the questions that have endured. Is the world safer when power is concentrated in a few states or broadly distributed (Waltz 1979)? How does the balance of power between states, or shifts thereof, affect the likelihood of war (Organski and Kugler 1980; Powell 1999, 2006)? Do international organizations allow states to gain benefits they would not receive from power politics alone (Keohane and Nye 1977)? Without good measures of power, we cannot provide good empirical answers to these fundamental questions. Consequently, the importance of measuring power to the study of international politics cannot be overstated.

Like many other important concepts in political science, power cannot be measured directly. Indeed, measurement problems in political science often entail the construction of proxies. Recent advances in computing and modeling have allowed political scientists to build sophisticated, data-driven proxies for variables as diverse as legislator ideology (Clinton, Jackman and Rivers 2004), judicial independence (Linzer and Staton 2014), and country regime types (Jackman and Treier 2008). But despite the centrality of power to many important hypotheses in international relations, its measurement has seen far less innovation.[1] In this article, we remedy this by devising a new, data-driven approach for measuring power. Specifically, we aim to learn what combination of observable material capability variables best predicts international dispute outcomes.

We are particularly interested in the crystallization of power that animates the bargaining model of war: the probability that one state will defeat another in case of militarized conflict, commonly denoted p. This outcome expectation is central to the standard bargaining model (Fearon 1995), where it serves as the operationalization of power when dismissing the mutual optimism hypothesis or when unearthing the commitment problem that arises due to shifts in power over time. The expected outcome of conflict also serves as the main concept of power

[1] A recent exception is Arena (2012).

in Slantchev's (2003) theory of war termination and in Powell's (2006) model of commitment problems. In other words, to study power—at least while motivated by the bargaining model—we must study what shapes dispute outcomes.

We focus on military capabilities as determinants of expected dispute outcomes. We do so for two reasons. First, there is a longstanding precedent of starting from the material foundations of power, a practice most commonly associated with historians and theorists in the realist camp (Morgenthau 1948; Taylor 1954; Carr 1964). And as Beckley (2010, 46) observes, even "[l]iberals and constructivists often conceptualize military power in material terms when refuting its causal significance." Second, and of more relevance for the current empirical literature, our focus on the contributions of material capabilities follows the example set by most existing efforts to measure power in the international sphere, starting with the work by Singer, Bremer and Stuckey (1972). Most current approaches use the Correlates of War Composite Index of National Capabilities (CINC) score, which combines material factors related to industrialization, wealth, population, and, of course, militarization.

Despite the innovations in measurement in various fields, political scientists have not reached a consensus on what makes for a good proxy, nor is there a common evaluatory metric. We argue for a predictive criterion: if the concept of interest is supposed to be associated with some observable outcome, then its proxy should predict the outcome well.[2] Simple as it may seem, this commitment to prediction highlights important issues. Like Ulysses or Goldilocks, the proxy maker must strike a delicate balance. She must learn from the data to construct the measure, else it will fail to capture important dimensions of the concept under study. *A priori* measures like summed rating scales suffer from this *underfitting* problem, as they fail to take advantage of the wealth of data scholars now possess. But the analyst who employs a data model for proxy construction faces pitfalls, too. She may misidentify chance features of her data as systematic, a problem called *overfitting*. A good proxy should fit the data well, but not so well that it fails to generalize. An underfit proxy will, of course, be a poor predictor, but so

[2]By prediction, we mean out-of-sample prediction, with data not used to construct the proxy itself.

will a data-driven proxy that maximizes in-sample fit at the expense of generalizability. Our predictive criterion balances these two considerations.

So too does our methodology. Supervised learning techniques, having been designed to navigate the straits between underfitting and overfitting, are ideal for data-driven proxy construction. Machine learning models are flexible enough to analyze relationships far more complex than possible in ordinary regression or measurement models, but they also guard against connecting the dots too aggressively or misinterpreting noise in the data as a complex relationship. To develop an optimal model for out-of-sample prediction, an analyst simply chooses appropriate tuning parameters, usually by a method like cross-validation that estimates prediction error (Efron and Gong 1983). Our approach mirrors that of Hill and Jones (2014), who use cross-validation to assess the relative predictive power of many variables all thought to affect the same outcome. Our focus, however, is on constructing variables rather than comparing them.

By the predictive criterion, a good measure of relative military power ought to predict dispute outcomes well. We show that the standard measure, the ratio of CINC scores, predicts militarized dispute outcomes terribly—only 1 percent better than random guessing. Our new proxy, the Dispute Outcome Expectations (DOE) score, is much better, providing a 20 percent predictive improvement. It is surprising that the standard measure does so poorly by this criterion, given its ubiquitousness. As we document below, dozens of recent publications in international relations use CINC-derived measures as proxies for power. Our use of modern machine learning tools allows us to yield a superior measure from the same data underlying the usual measure. In addition, the DOE score is interpretable as a probability, just like the bargaining concept of p that animates our approach.

In the course of developing the DOE score, we gain several broad insights about power. Most fundamentally, material capabilities indeed matter in shaping dispute outcomes, as we explain a substantial amount of variation with a small set of material variables. This basic result contrasts with previous studies finding no effect of material capabilities (Cannizzo 1980; Maoz

1983) and reinforces those that conclude capabilities affect victory (Bueno de Mesquita 1981; Stam 1996; Sullivan 2012). We go further to assess which of the CINC components matters most. Our results suggest that energy consumption is the strongest individual predictor of dispute outcomes. Surprisingly, military personnel and expenditures matter less on their own. However, it appears that the effect of these explicitly military components has evolved over the years, while energy consumption's effects have remained more static.

We then go on to demonstrate the DOE score's usefulness to international relations scholars. We replicate Reed et al.'s (2008) empirical test of Powell's (1996; 1999) model of the relationship between relative power, the distribution of benefits between states, and the likelihood of conflict. When we substitute DOE scores in for Reed et al.'s CINC-based proxy of p, the resulting model fits better and predicts better out-of-sample. More to the point, we extract an important new finding. Whereas Reed et al. conclude that the probability of conflict is sometimes greatest between states of equal power—namely, when the distribution of benefits is highly unequal—we find that this is never the case. The probability of war is always maximized when the state that is worse off under the status quo has a preponderance of power.

Interesting as these results are, it remains that most empirical practitioners only include a capability ratio as a control variable while modeling a wide variety of dependent variables. We take our replication further to see whether the DOE score would be helpful for these scholars as well. We reanalyze 18 such empirical models to see whether they fit better when we replace the standard proxy with DOE scores. Since these studies examine outcomes besides the one we use to construct our proxy, there is no guarantee that our new proxy will do better. Nonetheless, we yield an improvement in fit in 14 of the 18 cases. In the 14 improved cases, the DOE variables are always jointly significant, whereas the original CINC-based measures of power are jointly insignificant about half the time. Moreover, in two of these cases the main substantive hypothesis is no longer supported in the replicated model. We thereby show how using a poor proxy for relative power, even as a control variable, can lead scholars to understate the impact of military power and to reach conclusions not supported by the data.

The paper proceeds in five sections. In the first, we lay out our general argument about proxy construction and its application to the case of military power. Section 2 describes the data and methods we use to construct a new proxy for expected dispute outcomes. In Section 3, we discuss the advantages and disadvantages of our measure. Section 4 contains the results of our replications and advice for using the DOE score. The final section addresses next steps and concludes.

1 Proxies and Power

Prior to developing our proxy, we build on our discussion in the Introduction regarding the fundamental choices underlying our approach. Taken together, these points triangulate our general argument about what we mean by power and how to measure it.

1.1 Why Dispute Outcomes?

Like many, though not all, contemporary scholars, we orient our understanding of war in terms of the bargaining model. Bargaining models require disagreement points, and in the context of international crises, disagreement means conflict. The outcome of bargaining—whether an agreement is reached and, if so, which side it favors—depends on the likelihood of each potential outcome of fighting and the associated costs. As we discussed in the Introduction, the distribution of outcomes is usually pinned down through a single exogenous parameter, p, which captures the probability that one country defeats the other in case of war. In some of the most influential extant work on power (e.g., Powell 1999), p is its only analytic representative.

One of the bargaining model's most powerful features is its provision for p to influence *peaceful* outcomes, despite its inherently conflictual character. As Schelling (1966, 3) notes, "it is the *threat* of damage ... that can make someone yield or comply." In other words, the expected outcome of war sets the location of the bargaining range. This, in turn, is the reason that stronger states enjoy better peaceful settlements (Banks 1990). Consequently, we feel

comfortable taking victory in a dispute as an indicator of greater power even if the dispute did not proceed all the way to war.

Of course, the power to win hypothetical disputes is but one kind of power that states can exert over one another. There are other outcomes that might be relevant, and there may be ways that states influence one another that are not related to dispute outcomes. It is important to state explicitly our restriction in scope, but at the same time, this restriction is quite common for theorists and empiricists alike, especially those working within the bargaining paradigm.

1.2 Why Material?

Material capabilities are the starting point for much of what we know about power. Military historians have accrued impressive amounts of information on states' material holdings (e.g., Taylor 1954, Chapter 1), and realists have subsequently assigned materiel pride of place among explanators of power. As we noted above, material measures of power are also important to liberals and constructivists, if only for the sake of clearly ruling out realist accounts. The material conception of power has also served as a foil for those concerned with the particulars of force deployment or tactics more broadly (Biddle 2004).

More pragmatically, empirical scholars frequently use material capabilities, and particularly the ratio of CINC scores, as the data for their power proxies. Examining publications from 2005 to 2014 in five top journals for empirical international relations research,[3] we found at least 94 articles that control for the capability ratio or other proxies based on CINC scores. Though many of these articles' main models included other measures for channels of influence from one state to another like alliances or investment, it remains remarkable that such a broad swath of articles would include a measure of material capabilities. This is especially so because they cover a wide range of dependent variables, from conflict onset (easily the most common) to violations of international law to river treaty formation. Our material approach to power, then,

[3] *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *International Organization*, and *International Studies Quarterly*.

is of relevance to scholars across many areas of international relations.

By restricting our attention to the material dimensions of relative power, we also allow for an apples-to-apples comparison between our new measure and traditional CINC-based measures. Had we incorporated new covariates into our new measurement approach, it would be difficult to assess whether any gains (or losses) were due to the modeling strategy or the additional data. Doing so would also complicate the replication analysis we use to validate the new measure. Since all of the aforementioned studies include CINC scores or some function thereof in their regressions, we can proceed assuming that the material capability components are not endogenous, post-treatment, or otherwise unwise to include. If our measure included factors like alliance relationships or regime types, the universe of studies we could replicate to validate our approach would shrink considerably.

1.3 Why Prediction?

Our predictive criterion for measurement carries its own set of commitments. We adopt this criterion because it prioritizes models that generalize well—those that avoid the underfitting of *a priori* approaches and the overfitting of too-flexible approaches. The risk of overfitting is particularly high when there are too many degrees of freedom relative to the amount of data available. If the outcome of interest is only rarely observed, it might be hard to separate signal from noise. Similarly, overfitting is a concern if we are modeling the proxy as a function of many observable indicators, or we do not have the domain knowledge we would need to impose a specific functional form for the relationship between these indicators and the outcome of interest.

Situations like these are common in political science, including the current context. There are relatively few interstate disputes, and even fewer that involve just a single pair of states. Even if we restrict ourselves to the National Material Capabilities data, there is an abundance of variables: six capability components for each side of the dispute, along with the six annual shares associated with each raw component, for a total of 24. Incorporating time complicates

matters even further. As we do not have strong theory to guide us in choosing a functional form to relate capabilities to the probability of victory, we must instead take on our predictive approach.

The results from our predictive analysis thus map well to the analytic ideal of the probability of conflict in a hypothetical dispute between two states, which as we discuss above is an important variety of the more general concept of power. We should note that our predictive analysis does not imply that we are producing a leader's subjective belief that she will prevail in said hypothetical dispute. Instead, we produce the set of (objective) probabilities that best use the capabilities information at hand to predict the outcome. We should also note that we are not forecasting, as we might if we made more explicit use of training and test sets defined by time, but instead are using all the data at once and assessing prediction via cross-validation.

2 Building a Better Proxy for Relative Military Power

Our goal now is to squeeze as much predictive power as we can from data on states' material capabilities. When prediction is the goal, "black box" algorithmic techniques usually outpace standard regression models (Breiman 2001). So, to build our new measure, we augment traditional approaches with methods from machine learning.

2.1 Data

We combine the National Material Capabilities data (Singer, Bremer and Stuckey 1972) with information on the outcomes and participants of Militarized International Disputes between 1816 and 2007 (Palmer et al. 2015). Our data consist of $N = 1{,}740$ disputes, each between an "initiator," or Country A, and a "target," or Country B.[4] Every dispute outcome is either A Wins, B Wins, or Stalemate, denoted $Y_i \in \{A, B, \emptyset\}$. Most disputes end in a stalemate, and victory by the initiator is more than twice as likely as victory by the target, as shown in Table 1.

[4] See the Appendix for the data construction and coding specifics.

	Count	Proportion
A Wins	201	0.12
Stalemate	1460	0.84
B Wins	79	0.05

Table 1. Distribution of the three dispute outcomes.

We model dispute outcomes as a function of the participants' military capabilities. Our data source, the National Material Capabilities dataset, records annual observations of six characteristics of a country's military capability: military expenditures, military personnel, iron and steel production, primary energy consumption, total population, and urban population.[5] We also calculate each country's share of the global total of each component, giving us 12 variables per dispute participant. The matrix of predictors has 26 columns: the 24 individual capability characteristics of the initiator and target, the standard capability ratio, and the year the dispute began. Collect these predictors for the i^{th} dispute into the vector X_i.

2.2 A Metric for Predictive Power

As fortune plays a role in every military engagement, it is impossible to perfectly predict the outcome of every dispute. We therefore want a measure of predictive power that respects the probabilistic nature of militarized disputes. Classification metrics like the accuracy statistic, also known as the percentage correctly predicted, do not fit the bill. Instead, we employ the log loss, which is the negative of the average log-likelihood, as our metric for predictive power (Hastie, Tibshirani and Friedman 2009, 221). Let a *model* be a function $\hat{f}$ that maps from the dispute-level predictors X_i into the probability of each potential dispute outcome, $\hat{f}(X_i) = (\hat{f}_A(X_i), \hat{f}_B(X_i), \hat{f}_\emptyset(X_i))$. The "hat" on $\hat{f}$ emphasizes that the form of the function has been learned from the data, whether by estimating regression coefficients or by a more flexible

[5] About 17 percent of the disputes we observe contain at least one missing cell. We use multiple imputation to deal with missingness (Honaker and King 2010); see the Appendix for details.

predictive algorithm. The log loss of model $\hat{f}$ on the data (X, Y) is[6]

$$\ell(\hat{f}, X, Y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \{A,B,\emptyset\}} \mathbf{1}\{Y_i = y\} \log \hat{f}_y(X_i). \tag{1}$$

Smaller values of the log loss represent better predictive power, with the lower bound of 0 indicating perfect prediction.

We care mainly about the generalization error of our models—the expected quality of their predictions for new data that was not used to fit the models. To measure out-of-sample predictive power without losing data, we use K-fold cross-validation (Hastie, Tibshirani and Friedman 2009, 241–249).[7] Following standard practice, we set $K = 10$. Let $\text{CVL}(\hat{f})$ denote the 10-fold cross-validation estimate of the out-of-sample log loss. To ease interpretation, we compare models' log loss to that of a null model, whose predicted probabilities always equal the sample proportions of each outcome. The proportional reduction in cross-validation loss of the model $\hat{f}$ is

$$\text{PRL}(\hat{f}) = \frac{\text{CVL}(\hat{f}_{\text{null}}) - \text{CVL}(\hat{f})}{\text{CVL}(\hat{f}_{\text{null}})}. \tag{2}$$

The theoretical maximum, for a model that predicts perfectly, is 1. If a model predicts even worse than the null model—meaning it is worse than random guessing—its proportional reduction in loss is negative.

2.3 Modeling Dispute Outcomes

Our task now is twofold: to assess the predictive power of the capability ratio and, should we find it lacking (as we do), to build a better alternative.

We model dispute outcomes as a function of the capability ratio via ordered logistic regression (McKelvey and Zavoina 1975). To reduce skewness, we take the natural logarithm of the

[6] To avoid numerical problems, very low probabilities are trimmed at $\epsilon = 10^{-14}$.

[7] When dealing with models with tuning parameters that are themselves selected by cross-validation, we choose tuning parameters separately within each of the K iterations via another cross-validation loop (Varma and Simon 2006).

	Estimate	SE	Z	p
Capability Ratio (logged)	0.26	0.06	4.16	<0.01
Cutpoint: B Wins to Stalemate	−3.31	0.14		
Cutpoint: Stalemate to A Wins	1.84	0.09		

Table 2. Results of an ordered logistic regression of dispute outcomes on the capability ratio using the training data.

capability ratio. The parameter estimates from the capability ratio model on the full sample appear in Table 2. Although these results do not speak directly to the capability ratio's out-of-sample performance, they foreshadow why its predictive power is so limited. The coefficient on the capability ratio is statistically significant but small enough relative to the cutpoints that it always predicts a stalemate within the sample. This does not bode well for its out-of-sample performance.

We want a better model than what the capability ratio gives us, but we do not have a strong *a priori* sense of what the true relationship between material capabilities and dispute outcomes looks like. We use tools from machine learning that are designed to predict well without imposing much structure on the data, drawing a set of candidate models from the top-ten list by Wu et al. (2007) and from the best performers in the tests by Fernández-Delgado et al. (2014). After excluding those unsuited to our data, we end up with six predictive algorithms: C5.0, support vector machines, k-nearest neighbors, classification and regression trees, random forests, and ensembles of neural nets.[8] Each algorithm is widely used for prediction and can predict dispute outcome probabilities as a complex, potentially nonlinear function of the material capability components. As a compromise between these flexible "black box" models and the rigid capability ratio model, we also test ordered logistic regression models on the capability components.

In the spirit of flexibility, we try each model with different sets of predictors from the capability data. We examine four sets of variables: the raw capability components and the annual component shares, each with and without the year the dispute began. All of our models allow

[8] See the Appendix for full details of each method.

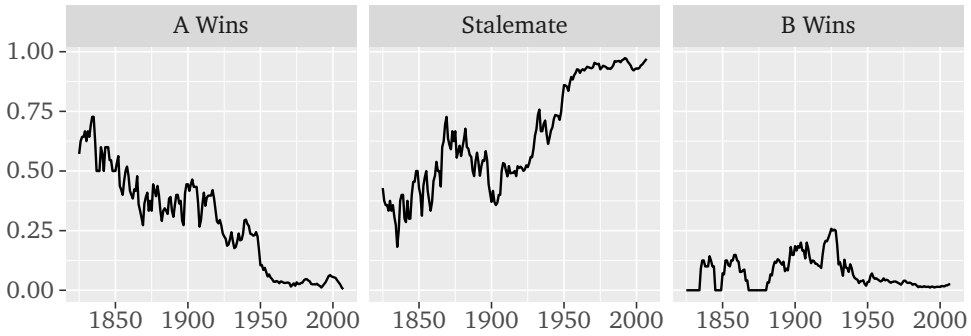

Figure 1. Distribution of dispute outcomes over time. Values are the proportion of disputes in the prior ten years ending in the given outcome.

for interactive relationships, so including the year of the dispute lets the effect of each capability component vary over time. With two sides per dispute and six capability variables per side, each model has 12 or 13 variables, depending on whether the year is included. To ensure that the models with the year included are not just picking up differences in the distribution of outcomes over time (see Figure 1), we also include an ordered logit of outcome on a third-order polynomial for year (Carter and Signorino 2010), a post-1945 dummy, and their interaction. All told, we have 31 candidate models: four sets of variables for each of our seven algorithms, plus the capability ratio model, the time trend model, and a null model used as a baseline.

We use cross-validation to estimate how well each of our candidate models predicts out of sample. The final problem, given those estimates, is to choose a model to construct an alternative to the capability ratio as a measure of expected dispute outcomes. It is tempting to simply pick the model with the lowest cross-validation loss. We can do even better at prediction, however, by taking a weighted average of all the models. We use the super learner algorithm (van der Laan, Polley and Hubbard 2007) to select the optimal model weights via cross-validation. Given a set of M candidate models $\hat{f}_1, \ldots, \hat{f}_M$, we select weights $\hat{w}_1, \ldots, \hat{w}_M$

to solve the constrained optimization problem

$$\begin{aligned} \min_{w_1,\ldots,w_M} \quad & \text{CVL}\left(\sum_{m=1}^{M} w_m \hat{f}_m\right) \\ \text{s.t.} \quad & w_1,\ldots,w_m \geq 0, \\ & w_1 + \ldots + w_m = 1, \end{aligned} \tag{3}$$

Our final model is the super learner, $\hat{f} = \sum_m \hat{w}_m \hat{f}_m$. Each individual model is a special case of the super learner, with full weight $\hat{w}_m = 1$ placed on a single $\hat{f}_m$. Hence, by the cross-validation criterion, we should prefer the super learner over any individual model.[9]

2.4 Cross-Validation Results

We now turn to the cross-validation results, which are summarized along with the super learner weights in Table 3. As the in-sample analysis hinted, the capability ratio is indeed a poor predictor of dispute outcomes. Its proportional reduction in loss is 0.01, which means its predicted probabilities are just 1 percent more accurate than the null model. This number is not encouraging, but what matters even more is whether we can do better. A glance at Table 3 confirms that we can: all but one of our 28 alternative models have greater predictive power than the capability ratio, many of them considerably better. With these results in hand, we feel comfortable dismissing the capability ratio as a suboptimal proxy for expected dispute outcomes.

As we expected, the super learner ensemble performs better than any of the candidate models from which it is constructed. The ensemble's proportional reduction in loss is about 23 percent, or four percentage points better than the best candidate model. Even after we apply a bias correction (see footnotes 7 and 9), the super learner's predictive power is still the

[9] As usual when selecting tuning parameters via cross-validation, the value of equation (3) is not an unbiased estimate of the generalization error of the super learner. Nested cross-validation is computationally infeasible for the super learner, so we calculate the bias correction recommended by Tibshirani and Tibshirani (2009) to estimate its generalization error.

Method	Data	Year	CV Loss	P.R.L.	Weight
Null Model	Intercept Only		0.54		<0.01
Ordered Logit	Capability Ratio		0.53	0.01	<0.01
Ordered Logit	Time Trend	✓	0.50	0.08	<0.01
Ordered Logit	Components		0.49	0.09	<0.01
Ordered Logit	Components	✓	0.48	0.10	<0.01
Ordered Logit	Proportions		0.51	0.04	<0.01
Ordered Logit	Proportions	✓	0.49	0.08	<0.01
C5.0	Components		0.53	0.01	0.01
C5.0	Components	✓	0.51	0.05	0.05
C5.0	Proportions		0.53	0.02	0.02
C5.0	Proportions	✓	0.52	0.04	<0.01
Support Vector Machine	Components		0.46	0.14	<0.01
Support Vector Machine	Components	✓	0.46	0.14	<0.01
Support Vector Machine	Proportions		0.49	0.09	<0.01
Support Vector Machine	Proportions	✓	0.48	0.11	<0.01
k-Nearest Neighbors	Components		0.47	0.12	<0.01
k-Nearest Neighbors	Components	✓	0.45	0.16	0.04
k-Nearest Neighbors	Proportions		0.51	0.05	<0.01
k-Nearest Neighbors	Proportions	✓	0.48	0.11	<0.01
CART	Components		0.52	0.02	<0.01
CART	Components	✓	0.44	0.18	0.18
CART	Proportions		0.54	−0.01	<0.01
CART	Proportions	✓	0.44	0.17	0.13
Random Forests	Components		0.49	0.08	0.02
Random Forests	Components	✓	0.49	0.09	0.20
Random Forests	Proportions		0.48	0.11	<0.01
Random Forests	Proportions	✓	0.48	0.11	0.01
Averaged Neural Nets	Components		0.44	0.19	0.08
Averaged Neural Nets	Components	✓	0.43	0.20	0.15
Averaged Neural Nets	Proportions		0.48	0.11	<0.01
Averaged Neural Nets	Proportions	✓	0.44	0.18	0.11
Super Learner			0.41	0.23	
(bias-corrected)			0.43	0.20	

Table 3. Summary of cross-validation results and super learner weights. All quantities represent the average across imputed datasets.

best among our models. Looking at the weights, what stands out is how few models are substantial components of the super learner: just five models have a weight of at least 5 percent. More generally, while models with lower generalization error tend to receive more weight, the relationship is by no means one-to-one. We see this because the ensemble prefers not only predictive power, but also diversity. Different classes of models have different blind spots; the more diverse the ensemble is, the more these blind spots are minimized. A model that looks bad on its own might still merit non-negligible weight in the optimal ensemble if it captures a slice of the data missed by the models that are best on their own.

For another illustration of the difference in predictive power between our model and the capability ratio, see the plots of out-of-fold predicted probabilities—the ones we use in cross-validation—in Figure 2. Under the capability ratio model, all but a handful of disputes are predicted to have an 80–90 percent chance of ending in stalemate. Seeing how narrow the capability ratio's predictive range is, it is little surprise that it barely does better than a null model at prediction. Conversely, the super learner makes much better use of the material capability data. Its predictive range is greater, which in turn allows it to achieve a stronger, though hardly perfect, relationship between predicted and observed outcomes.

2.5 Implications for International Relations

Our main focus is on developing a proxy for relative power that predicts the outcomes of militarized disputes, and predictive approaches like ours are not optimal for testing specific hypotheses (Shmueli 2010). Nonetheless, we can glean from our results a few important insights about the nature of the relationship between capabilities and power. The first is that there *is* a relationship—that variation in dispute outcomes is associated with variation in the disputants' raw capabilities. Our results therefore support the strand of literature finding that material capabilities influence dispute outcomes (Bueno de Mesquita 1981; Stam 1996; Sullivan 2012). Previous findings to the contrary (e.g., Cannizzo 1980; Maoz 1983) may simply reflect the inadequacy of CINC-based measures as a proxy for power.

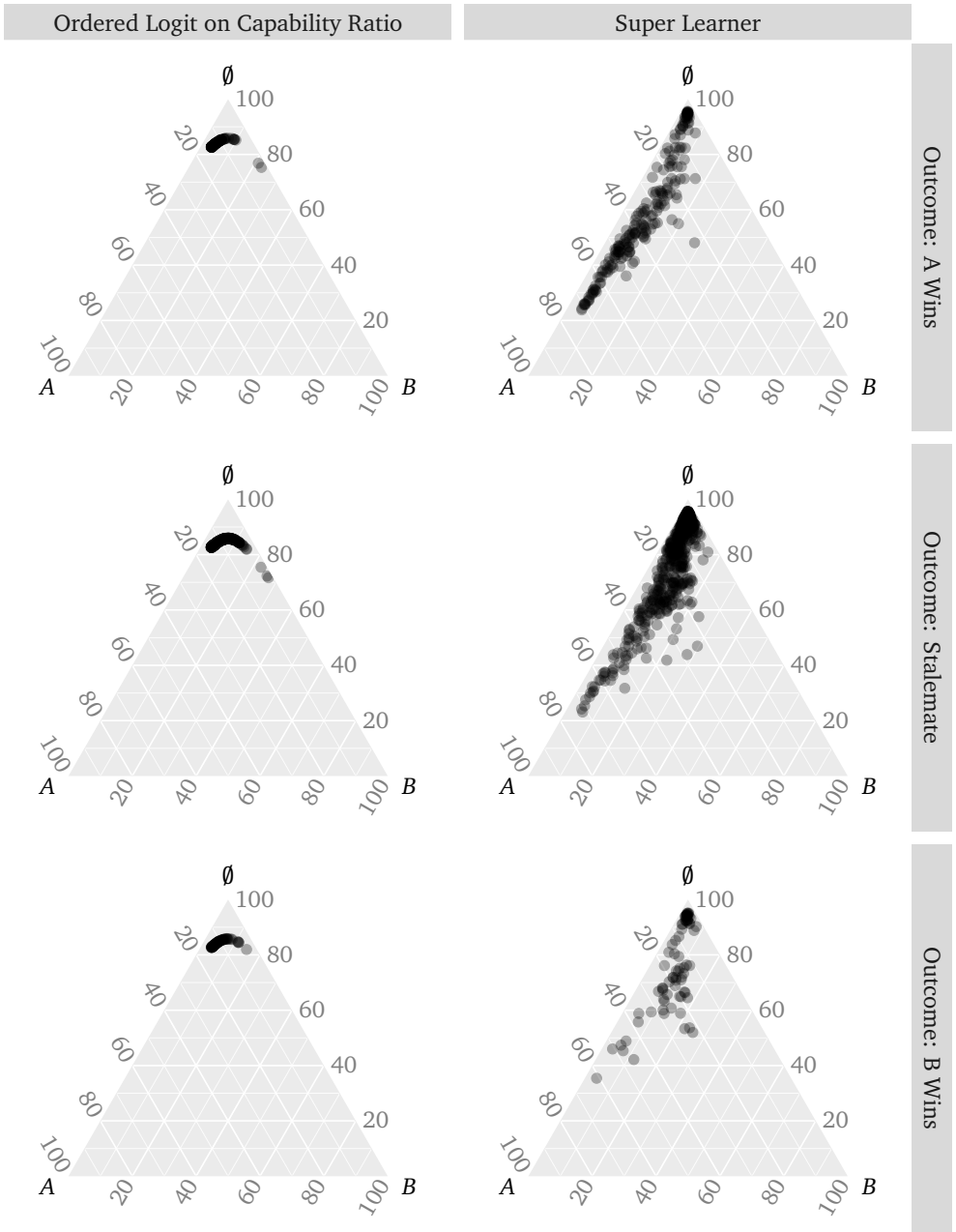

Figure 2. Ternary plots of out-of-fold predicted probabilities according to the capability ratio model and the super learner. Each predicted probability is calculated by fitting the model to 9/10 of the data, not including the observation in question—an approach that simulates true out-of-sample prediction.

	Increase in Loss	
Dropped Variable	With Year	Without Year
None	0.00%	2.37%
Iron and Steel Production	0.04%	2.62%
Military Expenditures	0.03%	5.24%
Military Personnel	0.33%	3.29%
Primary Energy Consumption	0.59%	2.17%
Total Population	−0.11%	2.69%
Urban Population	0.18%	3.07%

Table 4. Percentage increases in loss, relative to the full ensemble, due to removing each capability component from the analysis. The results "without year" come from running the super learner on only the 16 component models without the year variable.

But material power is not all that matters. Even after an intense, diverse predictive effort, we explain only 20 percent of the variation in dispute outcomes with material capability variables. To some extent this reflects the inherent unpredictability of military affairs; we would never expect to predict outcomes perfectly. Another potential source of error is that, depending on the extent of their aims, states may not fully deploy the capabilities they possess (Sullivan 2007). We suspect, however, that we could predict dispute outcomes even better by conditioning on more observable indicators. That is a task for future work, as the purpose of this paper is only to develop a proxy for the material components of relative power.

Though our approach has been more predictive than explanatory, we still would like some sense of which variables mattered the most in the predictive exercise. Had we focused on a single model alone, we might have had a simple answer to the question: for example, we might have looked at a random forest's variable importance statistics, or even just the coefficients in an ordered probit. Given our ensembled approach, however, the most straightforward way to proceed is to re-run the *entire* analysis many times, each time removing one of the predictors. If the removal of a given variable leads to a meaningful decrease in performance, we can infer that this variable is an important predictor. The leftmost column of Table 4 summarizes the results of this analysis. Here greater values indicate more loss of predictive power. Perhaps because the components are correlated with each other, the removal of any single component

does not change the results much. The greatest loss in predictive power comes from dropping primary energy consumption (PEC), an indicator of economic development and industrial capacity. It is interesting that PEC, a rough catch-all that correlates handily with any number of relevant constructs, winds up playing a stronger direct role than more explicitly militaristic factors like troops or military expenditures. However, recent work (e.g. Beckley 2010) suggests that economic development is a primary determinant of military effectiveness, while softer contextual factors like regime type, culture, or human capital play a far less vital role. The p we estimate here is entirely agnostic to functional form, but it seems that our flexible approach has allowed both direct-effort indicators like personnel to matter while also capturing the subtle interactive effects of a military effectiveness variable like economic development.

A second important finding is that the determinants of material power change over time. This conclusion may sound obvious, but it raises the question of why international relations scholars continue to use a proxy for power that assumes the relationship is unchanging. The simplest way to observe that time matters is to compare the predictive power of the models with and without the year variable: in 13 out of 14 cases, the model that includes time predicts better than its closest time-less counterpart.[10] The ensemble is not simply picking up the changing distribution of dispute outcomes over time: the model with a time trend alone has a mediocre PRL of 0.08, and it receives negligible weight.

For a look at how much the effect of each component changes over time, we turn back to the variable importance analysis, this time paying attention to the rightmost column of Table 4. Now each row represents a re-run of the analysis dropping a CINC component *and* the year variable. Again, we compare the results of each subsequent ensemble to the original, full ensemble. Doing so allows us to assess how much the predictive power of each CINC component varies with time; if a predictor matters little when dropped by itself but matters greatly when dropped along with time, then we might infer that its *dynamic* effects are important for the analysis. For example, the time variation is most pronounced for military expenditures, even

[10]The difference in log loss is statistically significant (paired $t = -3.1$, $p = 0.008$).

though it had a more pedestrian static effect (based on its lower score in the leftmost column). So, the returns to a dollar spent for military purposes have varied more over time than the returns to a soldier, or on increased industrial capacity, or increased population. This likely reflects changes in military technology and bureaucracy over time. As militaries have oscilatted between labor and capital intensity, so too have their requisite expense and the returns on investment (Howard 1976). The same goes for innovations in military bureaucracy.

3 The New Measure: Dispute Outcome Expectations

We use the super learner results to construct a new proxy for expected dispute outcomes—one that predicts actual dispute outcomes much more accurately than the capability ratio does. For any pair of countries at a particular point in time, whether or not they actually had a dispute with each other, we can use the super learner to ask, "Based on what we know about their material capabilities, how would a dispute between these countries be likely to end?" To construct the new proxy, we use the super learner to make predictions for every directed dyad–year in the international system between 1816 and 2007, the range of years covered by the National Material Capabilities data. We call the resulting dataset the Dispute Outcome Expectations data, or DOE. The DOE data contains predictions for more than 1.5 million directed dyad–years.[11] The canonical correlation between the DOE scores and the capability ratio is 0.44, so the measures are related but distinct.

The DOE scores are extrapolations. The overwhelming majority of dyad-years do not experience a dispute, and those that do are of course systematically different from those that do not. While we see the DOE scores as a significant advance in the state of the art of measuring power, we advise caution in their interpretation, particularly for dyads that would be unlikely to find themselves in a dispute. As the output of a model, DOE scores are estimates, and as

[11] About 19 percent of directed dyad–years contain missing values of at least one capability component. We average across imputations of the capabilities data to calculate the DOE scores for these cases. See the Appendix for details.

such they may be subject to selection bias. A promising direction for future work would be to develop data-driven proxies for power that preserve the flexibility of the super learner while more explicitly correcting for selection bias.

The DOE scores are naturally directed, since each dispute in our training data contains an initiating side and a target side. However, many analyses in the international conflict literature (e.g., of dispute occurrence) use undirected data. We calculate undirected DOE scores through a simple average of the directed values. For example, to calculate the probability that the United States would win a dispute against the United Kingdom in 1816, we average its estimated chances of victory as an initiator (36 percent) and as a target (11 percent) to yield 23.5 percent. If an analyst using the DOE data believed that the likely identity of an initiator in a hypothetical dispute were not a coin flip, she could take a different average of the directed scores to produce a more representative undirected score.

Perhaps counterintuitively, DOE scores should not be included as controls in regressions whose dependent variable is the outcome of a dispute or war. This may seem contradictory, given how much effort we have just spent showing that DOE scores are superior predictors of dispute outcomes. The reason they are superior is that, unlike the capability ratio, they are calibrated using real dispute data. But this in turn means that DOE scores would be endogenous in a regression whose dependent variable is dispute outcomes—i.e., the same data we used to construct the DOE scores.

4 Using the New Measure

A proxy's value ultimately depends on its usefulness in other applications. In this section, we demonstrate the DOE score's usefulness in two ways: first, through a detailed replication of a well-known test of the bargaining model; and second, through a replication of 18 recent studies that used other measures to proxy for relative power. We also provide some advice to practitioners on how to decide which measure(s) to include.

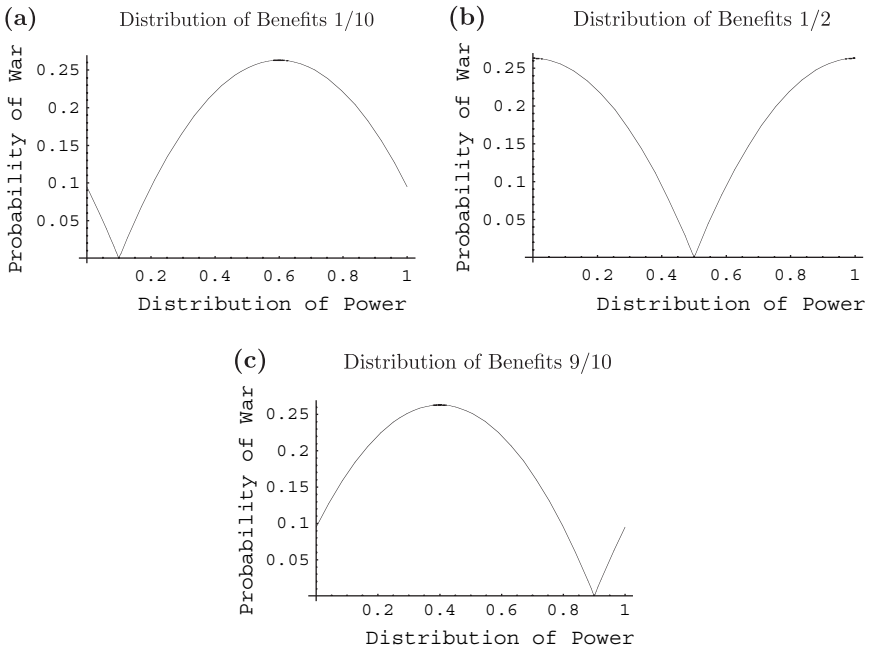

Figure 3. Reed et al.'s (2008) graphical summary of their main hypotheses.

4.1 Power, Benefits, and Conflict

The DOE score's greatest potential lies in its ability to enhance tests of the role of power in international relations. To that end, we replicate Reed et al. (2008), who study how the balance of power between two states affects the likelihood of conflict. They model the chance of interstate conflict as a function of two important parameters: the probability that one state would prevail over the other in a conflict, p, and the distribution of benefits between the two states, q. For example, q may capture where a border is drawn between two neighboring states. Motivated by Powell's (1996; 1999) theoretical model, they hypothesize that the effect of power depends on the status quo distribution of benefits. If benefits are distributed evenly between two states, conflict is most likely to break out if one state has a preponderance of power. Conversely, if the status quo disproportionately favors one state, conflict is most likely if there is a balance of power. Figure 3 summarizes these hypotheses.

Reed et al. test their theory by incorporating proxies of $|q-p|$ and $(q-p)^2$ (both lagged one year) into a model of dispute onset. Their measure of q, the distribution of benefits, is based on

	Reed et al. (2008)		DOE Replication	
Variable	Coefficient	S.E.	Coefficient	S.E.
Democracy	−0.011	0.004	−0.008	0.004
ln(Distance)	−0.205	0.003	−0.212	0.003
$\lvert q-p\rvert_{t-1}$	1.021	0.155	0.508	0.148
$(q-p)^2_{t-1}$	−0.617	0.196	0.426	0.170
Intercept	−1.580	0.026	−1.574	0.025
N	427,904		427,904	
AIC	12030.916		11808.073	
PRL	0.242		0.256	

Table 5. Replication of Table 1, Model 1 of Reed et al. (2008, 1213). The unit of analysis is the dyad-year, and the dependent variable is the onset of a militarized interstate dispute. The proportional reduction in loss over the null model comes from 100 repetitions of 10-fold cross-validation.

United Nations roll call votes. Like most of the conflict literature—not to mention this article—their measure of p, the dyadic balance of power, uses material capabilities. Specifically, their proxy for p is a normalized cousin of the capability ratio based on differences in CINC scores.[12] We replicate Reed et al.'s analysis, replacing the CINC-based measure of p with DOE scores while keeping all other covariates the same.[13] This is an ideal use case for DOE scores, since Reed et al., like us, draw from bargaining theory in treating power as the probability of success in an eventual conflict. The correlation between the original measure of $|q-p|$ and ours is 0.96. This is higher than the correlation between the capability ratio and DOE scores because we use the same measure of q.

Although our measure is similar, we yield substantively different results about the effect of the distribution of power and benefits on the likelihood of conflict. Table 5 summarizes the original analysis and our replication.[14] As we would expect, given the affinity between the DOE score and the bargaining model's concept of power, the model fit improves significantly

[12]For details, see footnote 11 of Reed et al. (2008, 1211).

[13]We report our replication of their Model 1. The results of our replication of their Model 2, which contains additional controls and peace-year splines, are substantively identical.

[14]We reconstruct Reed et al.'s measure of p with the latest National Material Capabilities data—the same we use to make DOE scores—so our sample size is slightly larger than in the original paper. The substantive and statistical significance of our estimates with the reconstructed measure, reported in the first column of Table 5, are the same as originally.

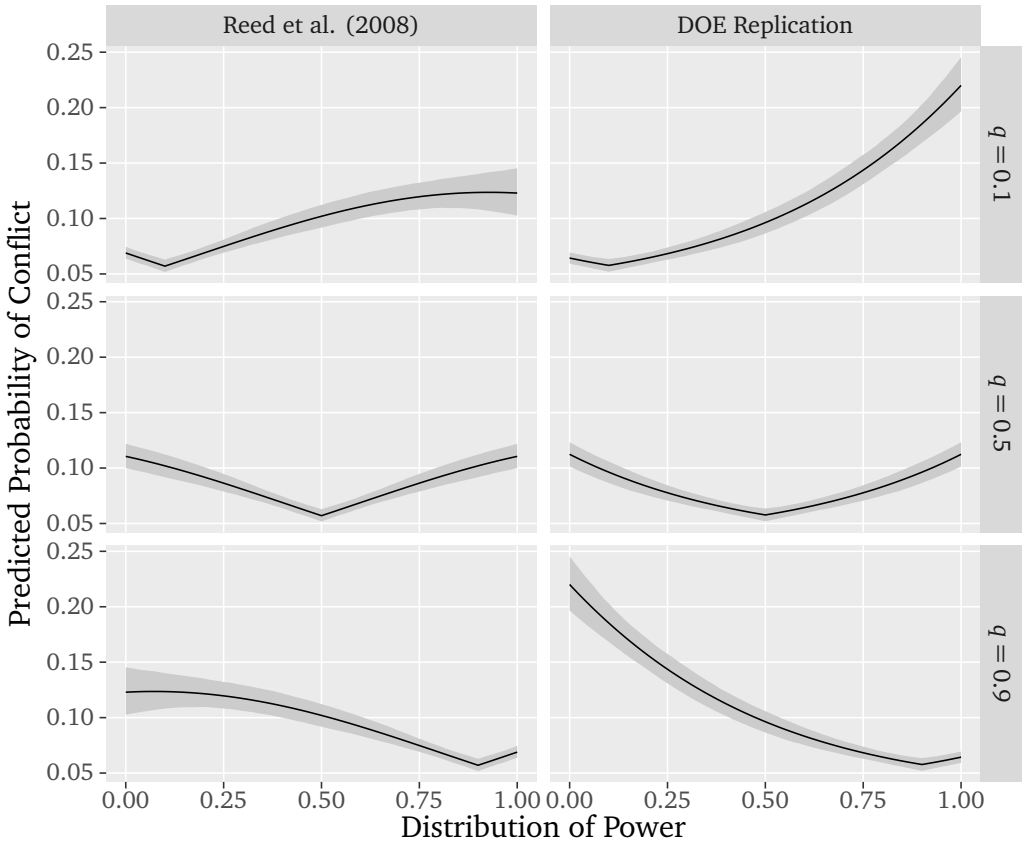

Figure 4. Replication of Figure 4 of Reed et al. (2008, 1213), which provides the predicted probability of MID onset as a function of the distribution of power (horizontal axis) and the distribution of benefits (vertical facets), holding democracy and distance at their minimal values. Standard errors obtained via a parametric bootstrap.

when we measure p with DOE scores instead of CINC scores. Using a Vuong (1989) test, we reject the null hypothesis of equal fit in favor of the DOE-based model fitting better ($Z = 7.80$, $p < 0.001$). The DOE model is also superior according to the AIC and cross-validation criteria. Accordingly, we feel comfortable making inferences from the replicated model.

The replicated model not only fits better, but also yields substantively different conclusions about the balance of power and war. Because of the nonlinear functional form of the model, we follow Reed et al. (2008) in leaning on graphical interpretations. Figure 4 plots the predicted probability of a dispute as a function of the balance of power and the distribution of benefits, according to the original model and our replication. Reed et al. (2008, 1212) cite their results, plotted in the first column, as "remarkable" support of the theoretical expectations reproduced

here in Figure 3. They find that neither power parity theory, which predicts conflict between evenly matched states, nor balance of power theory, which predicts conflict when one state holds a preponderance of power, holds unconditionally.[15] Instead, Reed et al. conclude that the power-conflict relationship resembles balance of power theory when benefits are evenly distributed ($q = 0.5$) and power parity theory when benefits are unequal ($q = 0.1$ or 0.9).

Our replication with DOE scores, plotted in the second column of Figure 4, leads us to overlapping but distinct conclusions. Like the original analysis, we find that the probability of conflict is always minimized when the distribution of benefits matches the distribution of power, or $q = p$. In addition, our results for the case when benefits are evenly matched are almost identical to Reed et al.'s. On the other hand, we never find support, even conditionally, for the power parity theory. Our model shows that the probability of conflict is always greatest when the difference between q and p is greatest—i.e., when the side with less benefits holds a preponderance of power. This finding runs contrary to both the theoretical expectations and the empirical findings of Reed et al., who claim that $p \approx 0.5$ is the most dangerous distribution of power when benefits are unequally distributed.

Though subtle, this different conclusion has important ramifications for the dyadic study of the balance of power. Indeed, our finding that the probability of conflict can *only* be maximized when the state with relatively fewer benefits has a perfect chance of winning runs counter to many studies in the dyadic tradition. Indeed, in the first page of the classic, foundational dyadic study, Bremer (1992, 309) notes that "a good deal of theoretical speculation and some empirical evidence suggest that war is more likely to occur between states that are ... roughly equal in power."[16] More forcefully, Lemke and Kugler (1996, 4) argue that "*parity* is the necessary condition for war." Later empirical extensions (e.g., Hegre 2008) have only been able to find qualified support[17] for such claims, and our results suggest that the DOE score might

[15] See Powell (1999, chapter 3) for further discussion of these schools of thought.

[16] It is worth noting that, based on binary "major power status" indicators, Bremer later goes on to rank power differences as relatively unimportant in both bivariate and multivariate analyses.

[17] As Hegre notes (586), "the analysis of power or size asymmetry and the risk of militarized interstate disputes shows that this relationship is far from straightforward."

have something to say in further evaluating such claims.

This also underscores the role of uncertainty in war onset. Of course, the variance of a binary outcome increases as it moves closer and closer to a 50/50 proposition. Thus, our result has the interpretation that, for any dyad with some given uncertainty about a hypothetical dispute outcome, there exists a nearby scenario with *less* uncertainty where war is *more* likely. In a naïve sense, this is a rather striking result. However, this only speaks to the importance of the introduction of q in Powell's original analysis and Reed et al.'s subsequent empirical investigation. Such a result might not obtain were states not so motivated to ensure that p and q aligned. So, while our original motivation was to provide a good empirical approximation of a parameter in the original, unmodified bargaining model of war, it remains that our improved measure can help us to appreciate the role of additional theoretical features, which in turn can improve the development of theory and empirics moving forward.

4.2 Capabilities as Control

In dyadic analyses of conflict, the capability ratio is often included as a control variable, but it remains important to use the best available proxy for power. Unless dyadic power relations have no effect on the outcome of interest (in which case proxies for power do not belong in the model), better proxies will capture more residual variation, resulting in greater model fit and more precise inferences. And if power is a confounding variable—i.e., power relations are correlated with both the key independent variable and the outcome—then the bias of the estimated relationship will be inversely related to the quality of the proxy. Reducing variance and bias are key concerns for any empirical analyst, so proxy quality matters.

To compare the performance of the DOE score as a control variable to that of the capability ratio, we replicate 18 recent analyses of conflict. In each replication, we rerun the main model with DOE scores in place of the capability ratio (or other CINC-derived proxy for relative power). Our main concern is fit: do the models with the DOE score capture more of the variation in the outcome of interest than those with the capability ratio? In 14 out of 18 cases, the

answer is yes, indicating that DOE scores make for a better control variable in typical statistical analyses of conflict.

We constructed the set of replications by looking for empirical analyses of dyad-years (directed or undirected) that included the capability ratio or another function of CINC scores as a covariate. Each study was published recently in a prominent political science or international relations journal.[18] We examined only studies with publicly available replication data. If we could not reproduce a study's main result or were unable to merge the DOE scores into the replication data (e.g., because of missing dyad-year identifiers), we excluded it from the analysis. We also excluded studies that employed duration models or selection models, due to conceptual and technical problems with assessing their out-of-sample performance. Lastly, we excluded studies in which our measure of expected dispute outcomes would be endogenous, namely those whose dependent variable was MID outcomes—the same data we used to construct the DOE scores—or a closely related quantity.[19] We were left with the 18 studies listed in Table 6.

For each analysis in our sample, we first identify the main statistical model reported in the paper, or at least a representative one.[20] We then estimate two models: the original model, and a replicated model where we replace any functions of CINC scores with their natural equivalents in DOE scores. For example, if the capability ratio is logged in the original model, we log the DOE scores in the replicated model. Our main measure of comparative model fit is the Vuong (1989) statistic for the test of the null hypothesis that the original and replicated models fit equally well.[21]

We are also interested in how DOE scores change our inferences about the main substantive variables or about the effect of power itself. We identify the main substantive hypothesis of

[18] For details, see footnote 3.

[19] The dependent variable of each study is listed in the Appendix. In most cases it is the initiation or onset of a dispute.

[20] When no main model is apparent, our heuristic is to pick one on the log-likelihood–sample size frontier. Details of the model chosen from each paper and the functions of CINC and DOE scores used are in the Appendix.

[21] We employ the standard Bayesian Information Criterion (Schwarz 1978) correction to the Vuong test statistic. We also measure model fit by the Akaike (1974) Information Criterion and repeated cross-validation; the results, which are reported in the Appendix, are essentially the same.

			Main Hyp.		Power Hyp.	
Replication	N	Vuong	p_{CINC}	p_{DOE}	p_{CINC}	p_{DOE}
Bennett (2006)	1,065,755	−13.21	✓	✓	✓	✓
Weeks (2012)	766,272	4.71	✓	✓	✓	✓
Jung (2014)	742,414	1.66	✓		✓	✓
Park and Colaresi (2014)	379,821	1.92	✓	✓		✓
Sobek, Abouharb and Ingram (2006)	183,227	3.43	✓	✓	✓	✓
Gartzke (2007)	171,509	4.09	✓	✓	✓	✓
Salehyan (2008*b*)	86,497	1.34	✓	✓	✓	✓
Fuhrmann and Sechser (2014)	85,306	1.40	✓	✓		✓
Arena and Palmer (2009)	54,403	2.76	✓		✓	✓
Owsiak (2012)	15,806	2.38	✓	✓	✓	✓
Zawahri and Mitchell (2011)	12,186	0.84	✓	✓	✓	✓
Salehyan (2008*a*)	10,197	1.49	✓	✓		✓
Fordham (2008)	7,788	−2.25	✓		✓	✓
Dreyer (2010)	5,316	2.52	✓	✓		✓
Huth, Croco and Appel (2012)	3,826	−0.69	✓		✓	✓
Uzonyi, Souva and Golder (2012)	1,667	1.57	✓	✓		✓
Weeks (2008)	1,582	1.26	✓	✓		✓
Morrow (2007)	864	−2.58	✓	✓	✓	✓

Table 6. Summary of results from the replication analysis. Positive values of the Vuong test statistic indicate that the model with DOE terms fits better than the model with CINC terms, and vice versa for negative values. The next two columns report whether $p < 0.05$ for the main substantive hypothesis test under each model; the final two report whether $p < 0.05$ for a test of the null hypothesis that all power variables have a coefficient of zero.

each study and perform the corresponding null hypothesis test on both the original model and the DOE score replication. The hypotheses tested are listed in the Appendix. To test for an effect of relative power, in the original models we test the null hypothesis that all CINC-derived terms have a coefficient of zero, and in the replication models we do the same with the DOE terms.

Table 6 summarizes the results of the replication analysis. The results further support our contention that DOE scores are superior to the capability ratio as a proxy for relative power. In a majority of the conflict studies we replicate, we explain more of the variation in the dependent variable when we replace the capability ratio with DOE scores as a control for power. According to the Vuong statistic, the DOE model fits better in 14 out of 18 cases; in six of these, the difference is statistically significant (the Vuong statistic exceeds 1.96). These results reinforce our confidence in the DOE score's quality as a proxy for relative power. They also affirm our conceptualization of relative power as the expected outcome of a dispute: by optimizing for dispute outcome prediction, we end up with a measure that is better for modeling a variety of other outcomes as well.

In most of the replications, the main substantive inference does not depend on the measure of relative power. That is not surprising, given that power is only a control variable in these studies. Focusing only on replications where the DOE score provided an improvement in fit and performance, two exceptions emerge. Interestingly, both are analyses of the international ramifications of domestic politics. The first is the study by Arena and Palmer (2009) examining the effects of major powers' government partisanship and economic conditions on their propensity to initiate disputes. Our replicated model with DOE scores both fits better and leads us not to reject the null hypothesis that government partisanship has zero effect (Wald $\chi^2 = 6.62$, df $= 8$, $p = 0.58$).[22] The second is Jung's (2014) analysis of diversionary conflict. The original study includes both the capability ratio and a CINC-based measure of rising powers; it interacts the latter with domestic unrest, a key independent variable of interest. When we replace the

[22]The null hypothesis is that government partisanship and its three interactions with economic variables have zero coefficient in both the mean and dispersion equations. See the Appendix for details.

capability ratio and the rising power measure with their DOE score equivalents, the resulting model fits better, and domestic unrest and its interaction with rising power are jointly insignificant (Wald $\chi^2 = 3.46$, df $= 2$, $p = 0.18$). By using a weak proxy for relative power, both of these analyses fail to pick up its confounding effects on the relationship of interest, leading them to overstate the effects of domestic pressures on international conflict.

The most striking results of the replication analysis come from the tests of the effects of power. In a third of the original studies, the relative power variables are statistically insignificant. One might conclude from these results that the importance of material power to international conflict is not robust. However, the DOE variables are jointly significant in every one of the replicated models. The insignificance of the capability ratio in many studies is not because power is unimportant, but because the capability ratio is such a poor proxy for power.

4.3 Advice to Practitioners

Seeing as neither the capability ratio nor DOE scores are uniformly better in typical applications, how should empirical scholars choose which one to include in their analysis? Our main recommendation is a theory-driven approach. When theory provides no guidance, we recommend either a data-driven approach or dropping capability measures altogether.

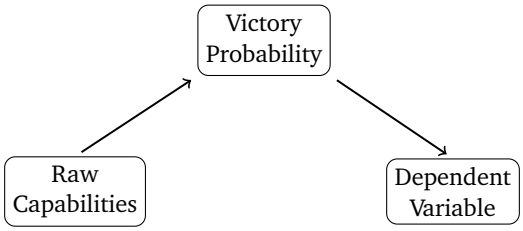

Figure 5. Raw capabilities only affect the outcome of interest through the probability of victory.

If theory suggests that material capabilities only affect the outcome of interest insofar as they shape the probability of victory, then DOE scores are the best measure to control for. Figure 5 contains a causal graph of this situation. One example of this scenario is the aforementioned test of Powell's (1996; 1999) theory of by Reed et al. (2008).

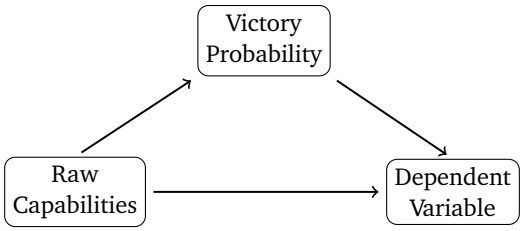

Figure 6. Raw capabilities affect the outcome of interest both directly and through expectations.

If material capabilities affect the outcome both directly and indirectly via victory probabilities, then it would be appropriate to control for both. Figure 6 illustrates this scenario. For example, imagine an empirical study of "sinking costs" via military mobilization in international crises (Fearon 1997). The initial movement of peaceful relations into a crisis, as well as early behavior at the bargaining table, might be shaped solely by states' expectations about dispute outcomes. But if states build up their military as a way to signal resolve, independently of the effect on likely outcomes, then raw capabilities matter too. When empirically modeling a theory like this, scholars should include both DOE scores and raw capability measures.

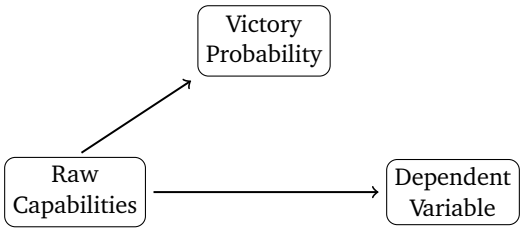

Figure 7. Raw capabilities directly affect the outcome of interest, but expectations do not.

The last possibility to consider is that expectations do not directly affect the outcome of interest. In this case, empirical models should only include raw capability measures, not DOE scores. The clearest example is when the dispute outcome itself is the dependent variable. Because DOE scores are calculated using the dispute outcome data, the DOE scores themselves are endogenous to observed outcomes, and thus should not be included as an independent variable when outcome is the dependent variable.[23]

When there is no specific theory about how material capabilities affect the outcome of in-

[23] In principle, this problem could be solved by only using data for years up to $t-1$ to calculate the DOE score for year t. Doing so is not feasible at present given the computational cost.

terest, we recommend a data-driven approach. The steps are the same ones we take above: determine a metric for model fit, run the model separately for each potential measure, and choose the best-fitting model. Alternatively, if your theory says nothing about the relationship between capabilities and the outcome of interest, it may be best not to include capability measures at all.

5 Conclusion

The DOE scores outperform the extant proxy—the CINC-based capability ratio—in a number of important ways. In pure terms, the DOE score more closely relates to what international relations scholars care about: the expected outcome of a dispute. On the practical side, our replications suggest that the DOE score is a better contributor to the usual battery of variables included in the ever-expanding universe of international relations regressions. Though it represents a massive improvement over the *status quo*, the DOE score could still be improved. We have only included those variables that could be extracted from the data used to construct the capability ratio. We did so consciously to demonstrate that our method could improve measures holding the covariates fixed. Having made our point, we look forward to future versions of DOE with new data is brought to bear on the problem. Since our underlying method uses well-programmed algorithms, anybody with a computer—and some patience!—could create a new version with new covariates.

On the methodological side, we believe that our data-driven approach to measurement will prove useful for those wishing to proxy for other quantities. All one needs is a set of predictor variables X and some outcome of interest Y—the procedure we provide to produce a mapping f from X to Y will work. Just as with introducing new covariates in any given application, future scholars can improve their proxies by including new models in the super learner. Our application tasked us to create a proxy of a probabilistic expectation, and similar applications provide a natural starting point for our method. Doing so, however, requires good

theory for just what we hope to predict with our abstractions. As such theories continue to develop, we hope political scientists across subfields will turn their attention to prediction and flexibility as they construct new measures and improve existing ones.

We would like to conclude with a still broader point. Breiman (2001) argues that statistical modelers fall into one of two cultures: data modelers, who interpret models' estimates after assessing overall quality via in-sample goodness of fit; and algorithmic modelers, who seek algorithms that predict responses as well as possible given some set of covariates. The method we advance is certainly algorithmic. Our decision to adopt algorithmic modeling based on prediction, however, was not culture-driven—it was purpose-driven (Clarke and Primo 2012). Most simply, prediction matters for measurement, so algorithmic tools should play a larger role. But as we show in the replication analysis, an algorithmically constructed proxy can be useful to include in traditional models. As new problems emerge and new solutions arise to solve them, we believe methodological pragmatism will be an important virtue. We neither expect nor encourage empirical political science to turn its focus from causal hypothesis testing to prediction. But good hypothesis testing depends on good measures, and sometimes the best way to build a measure is to assume the persona of the algorithmic modeler. By doing just that, this paper has developed one measure that improves on the previous state of the art along a number of dimensions.

References

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19(6):716–723.

Arena, Phil. 2012. "Measuring Military Capabilities." Blog post.
URL: http://fparena.blogspot.com/2012/11/once-more-on-military-capabilities.html

Arena, Philip and Glenn Palmer. 2009. "Politics or the Economy? Domestic Correlates of Dispute Involvement in Developed Democracies." *International Studies Quarterly* 53(4):955–975.

Banks, Jeffrey S. 1990. "Equilibrium Behavior in Crisis Bargaining Games." *American Journal of Political Science* 34(3):599–614.

Beckley, Michael. 2010. "Economic Development and Military Effectiveness." *Journal of Strategic Studies* 33(1):43–79.

Bennett, D. Scott. 2006. "Toward a Continuous Specification of the Democracy–Autocracy Connection." *International Studies Quarterly* 50(2):313–338.

Biddle, Stephen. 2004. *Military Power: Explaining Victory and Defeat in Modern Battle*. Princeton, NJ: Princeton University Press.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3):199–231.

Bremer, Stuart A. 1992. "Dangerous Dyads: Conditions Affecting the Likelihood of Interstate War, 1816–1965." *Journal of Conflict Resolution* 36(2):309–341.

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

Cannizzo, Cynthia A. 1980. The Costs of Combat: Death, Duration, and Defeat. In *The Correlates of War II: Testing Some Realpolitik Models*, ed. J. David Singer. pp. 233–257.

Carr, E.H. 1964. *The Twenty Years' Crisis: 1919–1939: An Introduction to the Study of International Relations*. New York: Harper and Row.

Carter, David B. and Curtis S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18(3):271–292.

Clarke, Kevin A. and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford, UK: Oxford University Press.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

Dreyer, David R. 2010. "Issue Conflict Accumulation and the Dynamics of Strategic Rivalry." *International Studies Quarterly* 54(3):779–795.

Efron, Bradley and Gail Gong. 1983. "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation." *The American Statistician* 37(1):36–48.

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

Fearon, James D. 1997. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *Journal of Conflict Resolution* 41(1):68–90.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro and Dinani Amorim. 2014. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* 15(1):3133–3181.

Fordham, Benjamin O. 2008. "Power or Plenty? Economic Interests, Security Concerns, and American Intervention." *International Studies Quarterly* 52(4):737–758.

Fuhrmann, Matthew and Todd S. Sechser. 2014. "Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Nuclear Deterrence." *American Journal of Political Science* 58(4):919–935.

Gartzke, Erik. 2007. "The Capitalist Peace." *American Journal of Political Science* 51(1):166–191.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Second ed. New York: Springer.

Hegre, Håvard. 2008. "Gravitating toward War Preponderance May Pacify, but Power Kills." *Journal of Conflict Resolution* 52(4):566–589.

Hill, Daniel W. and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3):661–687.

Honaker, James and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54(2):561–581.

Howard, Michael. 1976. *War in European History*. Oxford, UK: Oxford University Press.

Huth, Paul, Sarah Croco and Benjamin Appel. 2012. "Law and the Use of Force in World Politics: The Varied Effects of Law on the Exercise of Military Power in Territorial Disputes." *International Studies Quarterly* 56(1):17–31.

Jackman, Simon and Shawn Treier. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–217.

Jung, Sung Chul. 2014. "Foreign Targets and Diversionary Conflict." *International Studies Quarterly* 58(3):566–578.

Keohane, Robert O. and Joseph S. Nye. 1977. *Power and Interdependence: World Politics in Transition*. New York: Little Brown.

Lemke, Douglas and Jacek Kugler. 1996. The Evolution of the Power Transition Perspective. In *Parity and War: Evaluations and Extensions of the War Ledger*, ed. Jacek Kugler and Douglas Lemke. Ann Arbor, MI: University of Michigan Press pp. 3–34.

Linzer, Drew and Jeffrey K. Staton. 2014. "A Measurement Model for Synthesizing Multiple Comparative Indicators: The Case of Judicial Independence." Working paper.
URL: http://polisci.emory.edu/faculty/jkstato/resources/WorkingPapers/LS-scaling-140430.pdf

Maoz, Zeev. 1983. "Resolve, Capabilities, and the Outcomes of Interstate Disputes, 1816–1976." *Journal of Conflict Resolution* 27(2):195–229.

McKelvey, Richard D. and William Zavoina. 1975. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology* 4(1):103–120.

Morgenthau, Hans J. 1948. *Politics among Nations: The Struggle for Power and Peace*. New York: Alfred A. Knopf.

Morrow, James D. 2007. "When Do States Follow the Laws of War?" *American Political Science Review* 101(3):559–572.

Organski, A.F.K. and Jacek Kugler. 1980. *The War Ledger*. Chicago: University of Chicago Press.

Owsiak, Andrew P. 2012. "Signing Up for Peace: International Boundary Agreements, Democracy, and Militarized Interstate Conflict." *International Studies Quarterly* 56(1):51–66.

Palmer, Glenn, Vito D'Orazio, Michael Kenwick and Matthew Lane. 2015. "The MID4 dataset, 2002–2010: Procedures, Coding Rules and Description." *Conflict Management and Peace Science* 32(2):222–242.

Park, Johann and Michael Colaresi. 2014. "Safe Across the Border: The Continued Significance of the Democratic Peace When Controlling for Stable Borders." *International Studies Quarterly* 58(1):118–125.

Powell, Robert. 1996. "Stability and the Distribution of Power." *World Politics* 48(2):239–267.

Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics*. Princeton, NJ: Princeton University Press.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60(1):169–203.

Reed, William, David H. Clark, Timothy Nordstrom and Wonjae Hwang. 2008. "War, Power, and Bargaining." *Journal of Politics* 70(4):1203–1216.

Salehyan, Idean. 2008*a*. "No Shelter Here: Rebel Sanctuaries and International Conflict." *Journal of Politics* 70(1):54–66.

Salehyan, Idean. 2008*b*. "The Externalities of Civil Strife: Refugees as a Source of International Conflict." *American Journal of Political Science* 52(4):787–801.

Schelling, Thomas C. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6(2):461–464.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25(3):289–310.

Singer, J. David, Stuart Bremer and John Stuckey. 1972. Capability Distribution, Uncertainty, and Major Power War, 1820–1965. In *Peace, War, and Numbers*, ed. Bruce Russett. Beverley Hills, CA: Sage.

Slantchev, Branislav L. 2003. "The Principle of Convergence in Wartime Negotiations." *American Political Science Review* 97(4):621–632.

Sobek, David, M. Rodwan Abouharb and Christopher G. Ingram. 2006. "The Human Rights Peace: How the Respect for Human Rights at Home Leads to Peace Abroad." *Journal of Politics* 68(3):519–529.

Stam, Allan C. 1996. *Win, Lose, Or Draw: Domestic Politics and the Crucible of War*. University of Michigan Press.

Sullivan, Patricia L. 2007. "War Aims and War Outcomes: Why Powerful States Lose Limited Wars." *Journal of Conflict Resolution* 51(3):496–524.

Sullivan, Patricia L. 2012. *Who Wins? Predicting Strategic Success and Failure in Armed Conflict*. Oxford, UK: Oxford University Press.

Taylor, A.J.P. 1954. *The Struggle for Mastery in Europe, 1848–1918*. Oxford, UK: Clarendon Press.

Tibshirani, Ryan J. and Robert Tibshirani. 2009. "A Bias Correction for the Minimum Error Rate in Cross-Validation." *Annals of Applied Statistics* 3(2):822–829.

Uzonyi, Gary, Mark Souva and Sona N Golder. 2012. "Domestic Institutions and Credible Signals." *International Studies Quarterly* 56(4):765–776.

van der Laan, Mark J., Eric C. Polley and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6(1).

Varma, Sudhir and Richard Simon. 2006. "Bias in Error Estimation when Using Cross-Validation for Model Selection." *BMC Bioinformatics* 7(1):91.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2):307–333.

Waltz, Kenneth N. 1979. *Theory of International Politics*. Boston, MA: McGraw Hill.

Weeks, Jessica L. 2008. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization* 62(1):35–64.

Weeks, Jessica L. 2012. "Strongmen and Straw Men: Authoritarian Regimes and the Initiation of International Conflict." *American Political Science Review* 106(2):326–347.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg. 2007. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14(1):1–37.

Zawahri, Neda A. and Sara McLaughlin Mitchell. 2011. "Fragmented Governance of International Rivers: Negotiating Bilateral versus Multilateral Treaties." *International Studies Quarterly* 55(3):835–858.

A Appendix

A.1 National Material Capabilities Data

Our predictors are taken from the National Material Capabilities (v4.0) dataset from the Correlates of War project (Singer, Bremer and Stuckey 1972).[24] The dataset contains observations on six variables for 14,199 country-years from 1816 to 2007. For details on the variables and their measurement, see the NMC Codebook.[25] Table 7 lists the proportions of zeroes and missing values among each variable.

Component	Pr(Zero)	Pr(Missing)	θ
Iron and Steel Production	0.558	0.006	2^{-5}
Military Expenditures	0.034	0.139	2^{-7}
Military Personnel	0.066	0.027	2^{-1}
Primary Energy Consumption	0.097	0.030	2^{-3}
Total Population	0.000	0.002	2^{-7}
Urban Population	0.210	0.007	2^{-8}

Table 7. Proportions of zeroes and missing values in each National Military Capability component variable.

All six variables are right-skewed. Since five of the six variables are sometimes zero-valued (though all are non-negative), a logarithmic transformation is not appropriate. Instead, to correct for skewness, we apply an inverse hyperbolic sine transformation (Burbidge, Magee and Robb 1988) to each component:

$$h(x,\theta) = \sinh^{-1}(\theta x) = \log\left(\theta x + \sqrt{(\theta x)^2 + 1}\right). \tag{4}$$

We set the scale θ separately for each component variable with the aim of making the transformed variable approximately normally distributed. For each variable, we choose the value of $\theta \in \{2^d\}_{d=-10}^{10}$ that minimizes the Kolmogorov–Smirnov test statistic (Massey Jr. 1951) against a normal distribution with the same mean and variance. Table 7 gives the scale selected for each component. We use the transformed components in both the multiple imputation (see below) and the super learner training.

A.2 Militarized Interstate Dispute Data

Our sample and outcome variable are taken from the Militarized Interstate Disputes (v4.1) dataset from the Correlates of War project (Palmer et al. 2015).[26] The dataset records the

[24] Downloaded from http://correlatesofwar.org/data-sets/national-material-capabilities/nmc-v4-data/at_download/file.

[25] Available at http://correlatesofwar.org/data-sets/national-material-capabilities/nmc-codebook/at_download/file.

[26] Downloaded from http://correlatesofwar.org/data-sets/MIDs/mid-level/at_download/file.

participants and outcomes of interstate disputes from 1816 to 2010. To avoid the problem of aggregating capabilities across multiple states, we exclude disputes with more than one state on either side. We drop disputes that end in an outcome other than one side winning, one side yielding, or a stalemate;[27] we then collapse "A Wins" and "B Yields" into a single coding, and similarly for "B Wins" and "A Yields." Finally, since the capabilities data only run through 2007, we exclude disputes that end after 2007. In the end, we have $N = 1{,}740$ cases.

For each dispute in our dataset, we code the participating countries' capabilities using the values in the year the dispute began. About 17 percent of disputes have at least one missing capability component for at least one participant.

A.3 Multiple Imputation

As noted above, all of the National Material Capabilities variables contain some missing values. Following standard practice, we multiply impute the missing observations. We perform the imputations via the `Amelia` software package (Honaker, King and Blackwell 2011).

Rather than just impute the missing values in the final dataset of disputes, we impute the entire National Material Capabilities dataset. This allows us to fully exploit the dataset's time-series cross-sectional structure in the imputation process (Honaker and King 2010). We include in the imputation model a cubic polynomial for time, interacted with country dummy variables. As this results in an explosion in the number of parameters in the imputation model, we then impose a ridge prior equal to 0.1 percent of the observations in the dataset (see Section 4.7.1 of the `Amelia` package vignette). We enforce the constraint that every imputed value be non-negative. Finally, we impose an observation-level prior with mean zero and variance equal to that of the observed values of the corresponding component variable for every missing cell that meets the following criteria:

- There are no non-zero observed values in the time series preceding the cell
- The first observed value that comes after the cell is zero

So, for example, if a country's urban population is zero from 1816 to 1840, missing from 1841 to 1849, and zero in 1850, we would impose this form of prior on the 1841–1849 values. Diagnostic time series plots of observed versus imputed values within each data series, generated by the `tscsPlot()` function in `Amelia`, will be made available in the project's Dataverse.

The presence of missing data also complicates the calculations of country-by-country proportions of the total amount of each component by year. One option is to recompute the annual totals in each imputed dataset, so that the resulting data will be logically consistent—in particular, all proportions will sum to one. The drawback of this approach is that virtually every observation of the proportions will differ across the imputed datasets, even for countries with no missing data, since the annual totals will differ across imputations. An alternative approach is to compute the annual totals using only the observed values. The advantage is that non-missing observations will not vary across imputed datasets; the downside is that the proportions within each imputation will generally sum to more than one. For our purposes in this paper, we think it is preferable to reduce variation across imputations, even at the expense

[27] For details on other kinds of outcomes, see the MID Codebook.

of some internal consistency in the imputed datasets, so we take the latter approach: annual totals are the sums of only the observed values.

We impute $I = 10$ datasets of national capabilities according to the procedure laid out above, and we merge each with the training subset of our dispute data to yield I training data imputations. We run the super learner separately on each imputation, and our final model is an (unweighted) average of the I super learners.

After training is complete, we run into missing data problems once again when calculating DOE scores. To calculate predicted probabilities for dyads with missing values, we calculate a *new* set of $I = 10$ imputations of the capabilities data and take an (unweighted) average of our model's predictions across the imputations.

A.4 Super Learner Candidate Models

We use the R statistical environment (R Core Team 2015) for all data analysis. We fit, cross-validate, and calculate predictions from each candidate model through the `caret` package (Kuhn 2008). We then construct the super learner by solving (3) via base R's `constrOptim()` function for optimization with linear constraints. Candidate models were drawn from Wu et al. (2007) and Fernández-Delgado et al. (2014). Four of the algorithms named in Wu et al. (2007)—*k*-means, Apriori, expectation maximization, and PageRank—are not suited for the prediction task at hand. We also excluded AdaBoost due to long computation time and naive Bayes due to poor performance in initial tests. Further details about each candidate model are summarized below.

- Ordered Logit (McKelvey and Zavoina 1975)

  Package `MASS` (Venables and Ripley 2002)

  Tuning Parameters None

  Notes In the "Year" models, the year of the dispute is included directly and interacted with each capability variable

- C5.0 (Quinlan 2015)

  Package `C50` (Kuhn et al. 2015)

  Tuning Parameters

  - Number of boosting iterations (`trials`): selected via cross-validation from $\{1, 10, 20, 30, 40, 50\}$
  - Whether to decompose the tree into a rule-based classifier (`model`): selected via cross-validation
  - Whether to perform feature selection (`winnow`): selected via cross-validation

- Support Vector Machine (Cortes and Vapnik 1995)

  Package `kernlab` (Karatzoglou et al. 2004)

  Tuning Parameters

  - Kernel width (`sigma`): selected via cross-validation from $\{0.2, 0.4, 0.6, 0.8, 1\}$
  - Constraint violation cost (`C`): selected via cross-validation from $\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$

  Notes

  - Radial basis kernel
  - All predictors centered and scaled to have zero mean and unit variance

- k-Nearest Neighbors (Cover and Hart 1967)

  Package `caret` (Kuhn 2008)

  Tuning Parameters

  - Number of nearest neighbors to average (`k`): selected via cross-validation from $\{25, 50, \ldots, 250\}$

  Notes All predictors centered and scaled to have zero mean and unit variance

- CART (Breiman et al. 1984)

  Package `rpart` (Therneau, Atkinson and Ripley 2015)

  Tuning Parameters

  - Maximum tree depth (`maxdepth`): selected via cross-validation from $\{2, 3, \ldots, 9, 10\}$ (only up to 9 for models without year included)

- Random Forest (Breiman 2001)

  Package `randomForest` (Liaw and Wiener 2002)

  Tuning Parameters

  - Number of predictors randomly sampled at each split (`mtry`): selected via cross-validation from $\{2, 4, \ldots, 12\}$

  Notes 1,000 trees per fit

- Averaged Neural Nets (Ripley 1996)

  Package `nnet` (Venables and Ripley 2002), `caret` (Kuhn 2008)

  Tuning Parameters

  - Number of hidden layer units (`size`): selected via cross-validation from $\{1, 3, 5, 7, 9\}$
  - Weight decay parameter (`decay`): selected via cross-validation from $\{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$

  Notes Creates an ensemble of 10 neural nets, each initialized with different random number seeds

A.5 Replications

The following list contains basic information about each model in the replication study. We carry out logistic and probit regressions via `glm()` in base R (R Core Team 2015), multinomial logit via `multinom()` in the `nnet` package (Venables and Ripley 2002), ordered probit via `polr()` in the `MASS` package (Venables and Ripley 2002), and heteroskedastic probit via `hetglm()` in the `glmx` package (Zeileis, Koenker and Doebler 2013).

- Arena and Palmer (2009)

  Model Replicated Table 3

  Unit of Analysis Directed Dyads

  Dependent Variable MID initiation

  Estimator Heteroskedastic Probit

  CINC Terms CINC_A

  DOE Terms p_A, p_B

  Main Null Hypothesis

  $\beta[\text{Government}] = 0$
  $\beta[\Delta \text{ Unemployment * Government}] = 0$
  $\beta[\Delta \text{ Inflation * Government}] = 0$
  $\beta[\Delta \text{ Growth * Government}] = 0$
  (each in both the mean and dispersion equations)

  Notes CINC and DOE terms are included in both the mean and dispersion equations. Hypothesis tests use the nominal standard errors rather than the clustered standard errors reported in the paper.

- Bennett (2006)

  Model Replicated Table 1, Column 1

  Unit of Analysis Directed Dyads

  Dependent Variable MID initiation

  Estimator Logistic Regression

  CINC Terms CINC_A, CINC_B, CINC_{min} / CINC_{max}

  DOE Terms p_A, p_B, $|p_A - p_B|$

  Main Null Hypothesis

  $\beta[\text{Democracy-Initiator}] = 0$
  $\beta[\text{Democracy-Target}] = 0$
  $\beta[\text{Democracy-Initiator} \times \text{Similarity}] = 0$
  $\beta[\text{Democracy-Initiator} \times \text{Democracy-Target}] = 0$
  $\beta[(\text{Democracy-Initiator} \times \text{Democracy-Target})^2] = 0$

Replication	Model	AIC	CV P.R.L.	p_{main}	p_{power}
Arena and Palmer (2009)	CINC	1152	0.071†	5.51e-04	5.54e-27
Arena and Palmer (2009)	DOE	1055	0.141†	5.78e-01	8.17e-230
Bennett (2006)	CINC	29712	0.245†	1.40e-21	8.00e-97
Bennett (2006)	DOE	30912	0.215†	2.94e-21	1.98e-77
Dreyer (2010)	CINC	3676	0.239	4.17e-04	2.80e-01
Dreyer (2010)	DOE	3635	0.248	1.44e-03	2.69e-06
Fordham (2008)	CINC	537	0.275	1.17e-08	9.87e-16
Fordham (2008)	DOE	603	0.189	7.94e-02	1.48e-07
Fuhrmann and Sechser (2014)	CINC	2614	0.203	3.81e-05	5.13e-01
Fuhrmann and Sechser (2014)	DOE	2579	0.207	3.54e-04	4.34e-02
Gartzke (2007)	CINC	4284	0.442†	5.70e-10	2.50e-02
Gartzke (2007)	DOE	4162	0.458†	1.55e-09	6.34e-05
Huth, Croco and Appel (2012)	CINC	5938	0.053	3.47e-02	2.14e-04
Huth, Croco and Appel (2012)	DOE	5936	0.052	5.97e-02	4.32e-02
Jung (2014)	CINC	10665	0.350†	2.44e-04	8.79e-04
Jung (2014)	DOE	10611	0.353†	1.78e-01	2.26e-07
Morrow (2007)	CINC	1488	0.260	6.33e-30	1.44e-06
Morrow (2007)	DOE	1502	0.252	1.35e-28	1.09e-03
Owsiak (2012)	CINC	5805	0.117	2.71e-04	4.43e-05
Owsiak (2012)	DOE	5749	0.126	7.28e-14	2.40e-16
Park and Colaresi (2014)	CINC	10136	0.346†	6.50e-06	8.70e-02
Park and Colaresi (2014)	DOE	10114	0.347†	1.61e-06	3.60e-06
Salehyan (2008*b*)	CINC	8865	0.279	8.69e-06	8.86e-08
Salehyan (2008*b*)	DOE	8823	0.282	7.73e-05	1.71e-07
Salehyan (2008*a*)	CINC	3003	0.101	8.93e-12	4.15e-01
Salehyan (2008*a*)	DOE	2980	0.107	1.07e-08	4.64e-05
Sobek, Abouharb and Ingram (2006)	CINC	5165	0.347†	7.39e-24	6.03e-04
Sobek, Abouharb and Ingram (2006)	DOE	5066	0.359†	2.69e-17	4.18e-10
Uzonyi, Souva and Golder (2012)	CINC	2008	0.128	1.02e-03	7.22e-01
Uzonyi, Souva and Golder (2012)	DOE	1985	0.137	4.04e-04	7.57e-05
Weeks (2008)	CINC	1953	0.109	7.53e-04	4.05e-01
Weeks (2008)	DOE	1935	0.116	4.74e-04	3.31e-04
Weeks (2012)	CINC	15816	0.310†	2.33e-14	1.76e-08
Weeks (2012)	DOE	15572	0.321†	3.19e-12	3.86e-16
Zawahri and Mitchell (2011)	CINC	814	0.062	2.97e-09	2.49e-08
Zawahri and Mitchell (2011)	DOE	807	0.067	4.24e-10	7.69e-04

Table 8. Full results of replication analyses. Cross-validation results are from repeated 10-fold cross-validation; those marked † are repeated 10 times, all others 100 times.

- Dreyer (2010)

  Model Replicated Table 2, Model 2

  Unit of Analysis Undirected Dyads

  Dependent Variable MID occurrence

  Estimator Logistic Regression

  CINC Terms $\log(\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}})$

  DOE Terms $\log p_{\text{min}}$, $\log p_{\text{max}}$

  Main Null Hypothesis

  $$\beta[\text{Rapid issue accumulation}] = 0$$
  $$\beta[\text{Gradual issue accumulation}] = 0$$

- Fordham (2008)

  Model Replicated Table 2, third column (alliance onset with full set of controls)

  Unit of Analysis Undirected Dyads

  Dependent Variable Onset of alliance with U.S.

  Estimator Probit Regression

  CINC Terms $\log \text{CINC}_{\text{US}}$, $\log \text{CINC}_2$

  DOE Terms $\log p_{\text{US}}$, $\log p_2$

  Main Null Hypothesis $\beta[\text{Log of exports in previous year}] = 0$

- Fuhrmann and Sechser (2014)

  Model Replicated Table 2, Model 1

  Unit of Analysis Directed Dyads

  Dependent Variable MID initiation

  Estimator Probit Regression

  CINC Terms $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$

  DOE Terms p_A, p_B

  Main Null Hypothesis $\beta[\text{Defense pact with nuclear power}] = 0$

- Gartzke (2007)

  Model Replicated Table 1, Model 4

  Unit of Analysis Undirected Dyads

  Dependent Variable MID onset

  Estimator Logistic Regression

  CINC Terms $\log(\text{CINC}_{\text{max}} / \text{CINC}_{\text{min}})$

DOE Terms $\log p_{\min}$, $\log p_{\max}$

Main Null Hypothesis

$$\beta[\text{GDPPC (Low)}] = 0$$
$$\beta[\text{GDPPC} \times \text{Contig.}] = 0$$

- Huth, Croco and Appel (2012)

  Model Replicated Table 2

  Unit of Analysis Directed Dyads

  Dependent Variable Decision to challenge the status quo (keep status quo, negotiate, or threaten force)

  Estimator Multinomial Logistic Regression

  CINC Terms Average of A's respective shares of total dyadic military personnel, military expenditures, and military expenditures per soldier

  DOE Terms p_A, p_B

  Main Null Hypothesis $\beta[\text{Strong legal claims}] = 0$ in both the "Negotiations vs. threaten force" and "Status quo vs. force" equations

- Jung (2014)

  Model Replicated Table 1, Model 2

  Unit of Analysis Directed Dyads

  Dependent Variable MID initiation (threat to use force or greater)

  Estimator Logistic Regression

  CINC Terms $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$

  DOE Terms p_A, p_B

  Main Null Hypothesis

  $$\beta[\text{Unrest}] = 0$$
  $$\beta[\text{Unrest} \times \text{power}_{cinc}] = 0$$

  Notes In the DOE replication, power_{cinc} is replaced with a measure constructed analogously from DOE scores. It equals 1 for observations whose value of $(p_{B,t-1} - p_{B,t-5}) - (p_{A,t-1} - p_{A,t-5})$ is in the top quartile and 0 for all others.

- Morrow (2007)

  Model Replicated Table 1, first column (no weighting for data quality)

  Unit of Analysis Directed Dyads

  Dependent Variable Noncompliance with the laws of war (ordered: full compliance, high compliance, low compliance, noncompliance)

  Estimator Ordered Probit Regression

CINC Terms $\text{CINC}_A/(\text{CINC}_A + \text{CINC}_B)$, interaction with joint ratification and whether the violator lost

DOE Terms p_A, p_B, interactions of each with joint ratification and whether the violator lost

Main Null Hypothesis

$\beta[\text{Victim's Noncompliance}] = 0$

$\beta[\text{Clarity of Victim's Violations} \times \text{Victim's Noncompliance}] = 0$

$\beta[\text{Joint Ratification} \times \text{Victim's Noncompliance}] = 0$

$\beta[\text{Clarity of Victim's Violations} \times \text{Joint Ratification} \times \text{Victim's Noncompliance}] = 0$

$\beta[\text{Individual Violations} \times \text{Victim's Noncompliance}] = 0$

$\beta[\text{State Violations} \times \text{Victim's Noncompliance}] = 0$

Notes Capability ratio is "corrected for distance to the battlefield and aggregated across actors with a unified command." We drop the cases with coalitional actors in both models, hence the difference in sample size from the original article. No distance correction is applied to the DOE scores.

- Owsiak (2012)

  Model Replicated Table 3, Model 3

  Unit of Analysis Undirected Dyads

  Dependent Variable MID onset

  Estimator Logistic Regression

  CINC Terms $\log(\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}})$

  DOE Terms $\log p_{\text{min}}$, $\log p_{\text{max}}$

  Main Null Hypothesis $\beta[\text{Settled Borders}] = 0$

- Park and Colaresi (2014)

  Model Replicated Table 1, Model 4

  Unit of Analysis Undirected Dyads

  Dependent Variable MID onset

  Estimator Logistic Regression

  CINC Terms $\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}}$, interaction with contiguity

  DOE Terms $|p_A - p_B|$, interaction with contiguity

  Main Null Hypothesis $\beta[\text{Lowest Dem}] = 0$

- Salehyan (2008*a*)

  Model Replicated Table 1, Model 1

Unit of Analysis Undirected Dyads

Dependent Variable MID occurrence (category 4 or 5)

Estimator Logistic Regression

CINC Terms $\log(\text{CINC}_{\text{max}}/(\text{CINC}_{\text{max}}+\text{CINC}_{\text{min}}))$

DOE Terms $\log p_{\text{min}}$, $\log p_{\text{max}}$

Main Null Hypothesis $\beta[\text{External Base}]=0$

- Salehyan (2008*b*)

  Model Replicated Table 1, Model 1

  Unit of Analysis Directed Dyads

  Dependent Variable MID initiation

  Estimator Probit Regression

  CINC Terms $\text{CINC}_A/(\text{CINC}_A+\text{CINC}_B)$

  DOE Terms p_A, p_B

  Main Null Hypothesis

  $\beta[\text{Refugee Stock IN Initiator}]=0$
  $\beta[\text{Refugee Stock FROM Initiator}]=0$

- Sobek, Abouharb and Ingram (2006)

  Model Replicated Table 2, Index Model

  Unit of Analysis Undirected Dyads

  Dependent Variable MID onset

  Estimator Logistic Regression

  CINC Terms $(\text{CINC}_{\text{max}}-\text{CINC}_{\text{min}})/(\text{CINC}_{\text{max}}+\text{CINC}_{\text{min}})$

  DOE Terms p_{min}, p_{max}

  Main Null Hypothesis

  $\beta[\text{Physical Integrity Index}]=0$
  $\beta[\text{Empowerment Rights Index}]=0$

- Uzonyi, Souva and Golder (2012)

  Model Replicated Table 3, Model 3

  Unit of Analysis Directed Dyads

  Dependent Variable MID reciprocation

  Estimator Logistic Regression

  CINC Terms $\text{CINC}_A/(\text{CINC}_A+\text{CINC}_B)$

  DOE Terms p_A, p_B

Main Null Hypothesis $\beta[\text{ACC State A}] = 0$

- Weeks (2008)

  Model Replicated Table 4, Model 2

  Unit of Analysis Directed Dyads

  Dependent Variable MID reciprocation

  Estimator Logistic Regression

  CINC Terms $\text{CINC}_A/(\text{CINC}_A + \text{CINC}_B)$

  DOE Terms p_A, p_B

  Main Null Hypothesis

  $$\beta[\text{Single-Party}] = \beta[\text{Personalist}]$$
  $$\beta[\text{Military}] = \beta[\text{Personalist}]$$
  $$\beta[\text{Hybrid}] = \beta[\text{Personalist}]$$
  $$\beta[\text{Mixed Nondemocracy}] = \beta[\text{Personalist}]$$
  $$\beta[\text{Dynastic Monarchy}] = \beta[\text{Personalist}]$$
  $$\beta[\text{Nondynastic Monarchy}] = \beta[\text{Personalist}]$$
  $$\beta[\text{Nondemocratic Interregna}] = \beta[\text{Personalist}]$$
  $$\beta[\text{New Democracy}] = \beta[\text{Personalist}]$$

- Weeks (2012)

  Model Replicated Table 1, Model 2

  Unit of Analysis Directed Dyads

  Dependent Variable MID initiation

  Estimator Logistic Regression

  CINC Terms CINC_A, CINC_B, $\text{CINC}_A/(\text{CINC}_A + \text{CINC}_B)$

  DOE Terms p_A, p_B

  Main Null Hypothesis

  $$\beta[\text{Junta}] = \beta[\text{Machine}]$$
  $$\beta[\text{Boss}] = \beta[\text{Machine}]$$
  $$\beta[\text{Strongman}] = \beta[\text{Machine}]$$
  $$\beta[\text{Other Nondemocracies}] = \beta[\text{Machine}]$$

- Zawahri and Mitchell (2011)

  Model Replicated Table 2, Model 1

  Unit of Analysis Directed Dyads

  Dependent Variable River treaty formation

Estimator Logistic Regression

CINC Terms CINC_A, CINC_B

DOE Terms p_A, p_B

Main Null Hypothesis

$$\beta[\%\text{ Lowest area in basin}] = 0$$
$$\beta[\text{Lowest water dependence}] = 0$$
$$\beta[\text{Lowest avg. precipitation}] = 0$$

Notes Dyads are directed, but A is the upstream state in a river basin rather than the (prospective) initiator of conflict, so we use the undirected form of the DOE scores.

Additional References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone and R. A. Olshen. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.

Burbidge, John B., Lonnie Magee and A. Leslie Robb. 1988. "Alternative Transformations to Handle Extreme Values of the Dependent Variable." *Journal of the American Statistical Association* 83(401):123.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20(3):273–297.

Cover, T. M. and P. E. Hart. 1967. "Nearest Neighbor Pattern Classification." *IEEE Transactions on Information Theory* 13(1):21–27.

Honaker, James and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54(2):561–581.

Honaker, James, Gary King and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(7):1–47.
URL: http://www.jstatsoft.org/v45/i07/

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik and Achim Zeileis. 2004. "kernlab – An S4 Package for Kernel Methods in R." *Journal of Statistical Software* 11(9):1–20.
URL: http://www.jstatsoft.org/v11/i09/

Kuhn, Max. 2008. "Building Predictive Models in R Using the `caret` Package." *Journal of Statistical Software* 28(5):1–26.
URL: http://www.jstatsoft.org/v28/i05

Kuhn, Max, Steve Weston, Nathan Coulter, Mark Culp and Ross Quinlan. 2015. *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.0-24.
URL: http://CRAN.R-project.org/package=C50

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.
URL: http://CRAN.R-project.org/doc/Rnews/

Massey Jr., Frank J. 1951. "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association* 46(253):68–78.

McKelvey, Richard D. and William Zavoina. 1975. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology* 4(1):103–120.

Palmer, Glenn, Vito D'Orazio, Michael Kenwick and Matthew Lane. 2015. "The MID4 dataset, 2002–2010: Procedures, Coding Rules and Description." *Conflict Management and Peace Science* 32(2):222–242.

Quinlan, Ross. 2015. "Data Mining Tools See5 and C5.0." RuleQuest website.
URL: https://www.rulequest.com/see5-info.html

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: http://www.R-project.org/

Ripley, Brian D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Singer, J. David, Stuart Bremer and John Stuckey. 1972. Capability Distribution, Uncertainty, and Major Power War, 1820–1965. In *Peace, War, and Numbers*, ed. Bruce Russett. Beverley Hills, CA: Sage.

Therneau, Terry, Beth Atkinson and Brian Ripley. 2015. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-10.
URL: http://CRAN.R-project.org/package=rpart

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth ed. New York: Springer. ISBN 0-387-95457-0.
URL: http://www.stats.ox.ac.uk/pub/MASS4

Zeileis, Achim, Roger Koenker and Philipp Doebler. 2013. *glmx: Generalized Linear Models Extended*. R package version 0.1-0.
URL: http://CRAN.R-project.org/package=glmx