AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Prediction, Proxies, and Power

Robert J. Carroll Florida State University
Brenton Kenkel Vanderbilt University

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 (DOE) 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 and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/FPYKTP.

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), in which 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 in Slantchev's (2003) theory of war termination and in Powell's (2006) model of commitment problems.

Robert J. Carroll is Assistant Professor, Department of Political Science, Florida State University, 600 W. College Avenue, Tallahassee, FL 32306 (rjcarroll@fsu.edu). Brenton Kenkel is Assistant Professor, Department of Political Science, Vanderbilt University, 324 Commons Center, PMB 0505, Nashville, TN 37203 (brenton.kenkel@vanderbilt.edu).

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. We thank the authors listed in Table 6 for making their replication data publicly available. James Martherus and Bryan Rooney provided excellent research assistance. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University.

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

American Journal of Political Science, Vol. 63, No. 3, July 2019, Pp. 577–593

DOI: 10.1111/ajps.12442

In other words, to study power—at least while motivated by the bargaining model—we must study what shapes dispute outcomes.

We focus on material capabilities as determinants of expected dispute outcomes. In doing so, we follow 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, or 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 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 dispute outcomes only 1% better than random guessing. It is surprising that the capability ratio does so poorly, given its ubiquitousness. Our new proxy, the Dispute Outcome Expectations (DOE) score, is much better, providing a 20% predictive improvement. 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 concluding capabilities affect victory (Bueno de Mesquita 1981; Stam 1996; Sullivan 2012). Digging further, 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, whereas 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 and colleagues' (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 for Reed and colleagues' CINC-based proxy of p, the resulting model fits better and predicts better out-of-sample. 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.

We take our replication further to see whether the DOE score would be helpful when relative power is simply a control variable, as is typical in empirical models

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

of international politics. We reanalyze 18 such 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 these 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 article 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. The second section describes the data and methods we use to construct a new proxy for expected dispute outcomes. In the third section, we discuss the advantages and disadvantages of our measure. The fourth section contains the results of our replications and advice for using the DOE score. The final section addresses next steps and concludes.

Proxies and Power

Before developing our proxy, we build on our discussion in the introduction regarding the fundamental choices underlying our approach—what we mean by power and how to measure it.

Why Dispute Outcomes?

Like many, though not all, contemporary scholars, we orient our understanding of war in terms of the bargaining model. 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. In bargaining models, the distribution of outcomes is usually pinned down through a single exogenous parameter, p, the probability that one country defeats the other in case of war.

One of the bargaining model's most powerful features is its provision for p to influence *peaceful* outcomes. 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.

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, chap. 1), and realists have subsequently assigned material pride of place among explanators of power. Material measures of power are also important to liberals and constructivists, if only for the sake of clearly ruling out realist accounts (Beckley 2010, 46).

Empirical analyses 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 (the most common) to violations of international law to river treaty formation. Our material approach to power, then, 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

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

aforementioned studies include CINC scores or some function thereof in their regressions, we can proceed assuming that the material capability components are not endogenous, posttreatment, 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.

Why Prediction?

We adopt a predictive criterion for measurement 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. When the outcome of interest is only rarely observed, it is 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 instead take a predictive approach.

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.

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. We augment ordinary statistical approaches with "black box" machine learning tools, which usually outperform traditional models at prediction (Breiman 2001).

Table 1 Distribution of the Three Dispute Outcomes

	Count	Proportion
A Wins	201	0.12
Stalemate	1,460	0.84
B Wins	79	0.05

Data

We combine the National Material Capabilities data (Singer, Bremer, and Stuckey 1972) with information on the outcomes and participants of militarized international disputes (MIDs) 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.

We model dispute outcomes as a function of the participants' military capabilities. Our data source, the National Material Capabilities data set, records annual observations of six country characteristics: military expenditures, military personnel, iron and steel production, primary energy consumption, total population, and urban population.[5] We also calculate each country's annual share of the global total of each component, giving us 12 variables per dispute participant. Together with the standard capability ratio and the year the dispute began, these give us 26 predictors total, which we denote by the vector X_i.

A Metric for Predictive Power

We use a continuous measure to gauge predictive power, in contrast with more discrete metrics like the percentage correctly predicted. Specifically, we use the log loss, which is closely related to the log-likelihood from

[4] See the supporting information (SI) for the data construction and coding specifics.

[5] About 17% 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 SI for details.

traditional statistics (Hastie, Tibshirani, and Friedman 2009, 221). Let a *model* be a function $\hat{f}$, typically learned from the data, 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 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 were not used in estimation. To estimate generalization error without losing data, we use K-fold cross-validation (Hastie, Tibshirani, and Friedman 2009, 241–49). Following standard practice, we set $K = 10$. Let $\mathrm{CVL}(\hat{f})$ denote the 10-fold cross-validation estimate of the out-of-sample log loss. To ease interpretation, we compare this to 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

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

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

Modeling Dispute Outcomes

Our task now is 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 natural logarithm of the capability ratio via ordered logistic regression (McKelvey and Zavoina 1975), reported in Table 2. 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 speak directly to the generalization error of the capability ratio, but it illustrates the weakness of its predictive power.

To improve upon the capability ratio, we use tools from machine learning that are designed to predict well without imposing much structure on the data. We use learning methods from the top 10 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.[7] Each algorithm can predict dispute outcome probabilities as a nonlinear function of the material capability components. As a compromise between these "black box" models and the rigid capability ratio model, we also test ordered logistic regression models on the capability components.

TABLE 2 Results of an Ordered Logistic Regression of Dispute Outcomes on the Capability Ratio using the Training Data

	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		

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 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.

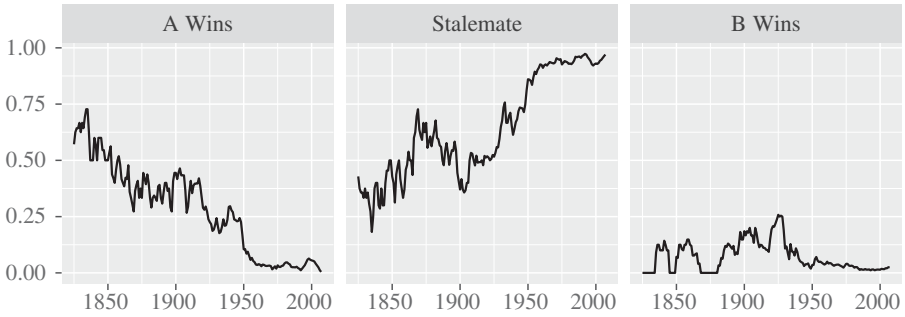
Instead of simply selecting the best performer among the candidate models, we use the super learner algorithm (van der Laan, Polley, and Hubbard 2007) to construct an optimal weighted ensemble.[8] To ensure that the

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

[7] See the SI for full details of each method.

[8] For mathematical details, see the SI.

FIGURE 1 Distribution of Dispute Outcomes Over Time



Note: Values are the proportion of disputes in the prior 10 years ending in the given outcome.

generalization error of the final ensemble is indeed lower than that of any given component model, we estimate its generalization error using the bias correction recommended by Tibshirani and Tibshirani (2009).

Cross-Validation Results

We now turn to the cross-validation results, which are summarized along with the super learner weights in Table 3. The capability ratio is indeed a poor predictor of dispute outcomes; its proportional reduction in loss is 0.01, meaning it is just 1% better than always guessing the average. 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.

The super learner has lower generalization error than any of the candidate models from which it is constructed. Looking at the ensemble weights, what stands out is how few models are substantial components of the super learner; just five have a weight of at least 5%. More generally, although 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% chance of ending in a 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.

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.

But material power is not all that matters. Using material variables as efficiently as possible, we still explain only 20% of the variation in dispute outcomes. To some extent, this reflects the inherent unpredictability of military affairs. 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

TABLE 3 Summary of Cross-Validation Results and Super Learner Weights

Method	Data	Year	CV Loss	PRL	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	0.04
C5.0	Proportions		0.52	0.02	0.01
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	0.13	<0.01
k-Nearest Neighbors	Components	√	0.45	0.16	0.01
k-Nearest Neighbors	Proportions		0.51	0.04	<0.01
k-Nearest Neighbors	Proportions	√	0.48	0.10	<0.01
CART	Components		0.52	0.04	<0.01
CART	Components	√	0.44	0.18	0.19
CART	Proportions		0.53	0.01	<0.01
CART	Proportions	√	0.44	0.18	0.12
Random Forests	Components		0.50	0.07	0.02
Random Forests	Components	√	0.50	0.08	0.19
Random Forests	Proportions		0.47	0.12	0.01
Random Forests	Proportions	√	0.47	0.12	<0.01
Averaged Neural Nets	Components		0.43	0.19	0.12
Averaged Neural Nets	Components	√	0.43	0.20	0.16
Averaged Neural Nets	Proportions		0.48	0.11	<0.01
Averaged Neural Nets	Proportions	√	0.44	0.18	0.12
Super Learner			0.41	0.23	
(Bias-Corrected)			0.43	0.20	

Note: All quantities represent the average across imputed data sets.

outcomes even better by conditioning on more observable indicators. That is a task for future work, as the purpose of this artile is only to develop a proxy for the material components of relative power.

To gauge the importance of individual variables, we rerun the predictive analysis numerous times, each time removing a single predictor. The most important predictors are those whose removal leads to the greatest decreases in predictive power. The leftmost column of Table 4 summarizes the results of this analysis; 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, an indicator of economic development and industrial capacity. It is surprising that this indicator is more predictive on its own than explicitly militaristic factors like troops and military expenditures. This finding mirrors recent work suggesting that economic development is a primary determinant of military effectiveness (Beckley 2010).

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 measure that assumes the relationship is unchanging. The simplest

FIGURE 2 Ternary Plots of Out-of-Fold Predicted Probabilities According to the Capability Ratio Model and the Super Learner

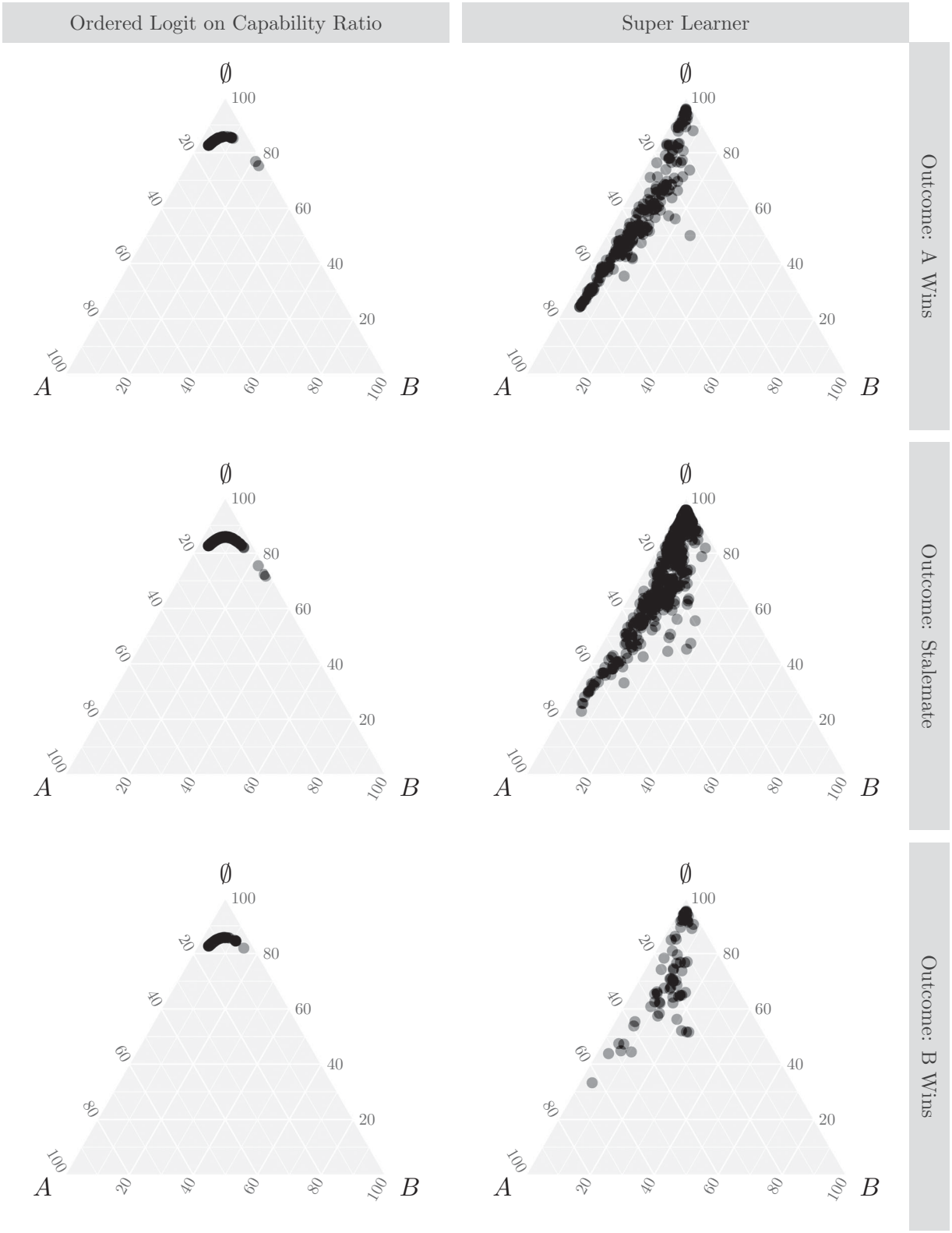


TABLE 4 Importance of Component Variables

Dropped Variable	Increase in Loss	
	With Year	Without Year
None	0.00%	2.16%
Iron and Steel Production	0.33	2.98
Military Expenditures	0.17	5.44
Military Personnel	0.56	3.48
Primary Energy Consumption	0.60	2.43
Total Population	0.10	2.87
Urban Population	0.36	3.35

Note: Each entry is the percentage increase in loss, relative to the full ensemble, due to removing the given 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.

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.[9] We are not simply picking up the changing distribution of dispute outcomes over time; the model with only a time trend performs less than half as well as the full ensemble.

To gauge the extent to which the effect of each variable changes over time, the rightmost column of Table 4 reports the increase in generalization error when the given variable and the year indicator are both removed from estimation. 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. The time variation is most pronounced for military expenditures, likely reflecting changes in military technology and bureaucracy over time. As militaries have oscillated between labor and capital intensity, so too have their requisite expense and the returns on investment (Howard 1976).

The New Measure: Dispute Outcome Expectations

We use our ensemble model 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 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.[10] We call the resulting data set the Dispute Outcome Expectations data (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 systematically differ from those that do not. Although 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 such they may be subject to selection bias. A promising direction for future work would be to develop data-driven proxies for relative power that preserve the flexibility of the super learner while more explicitly correcting for selection bias.

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—that is, the same data we used to construct the DOE scores.

Using the New Measure

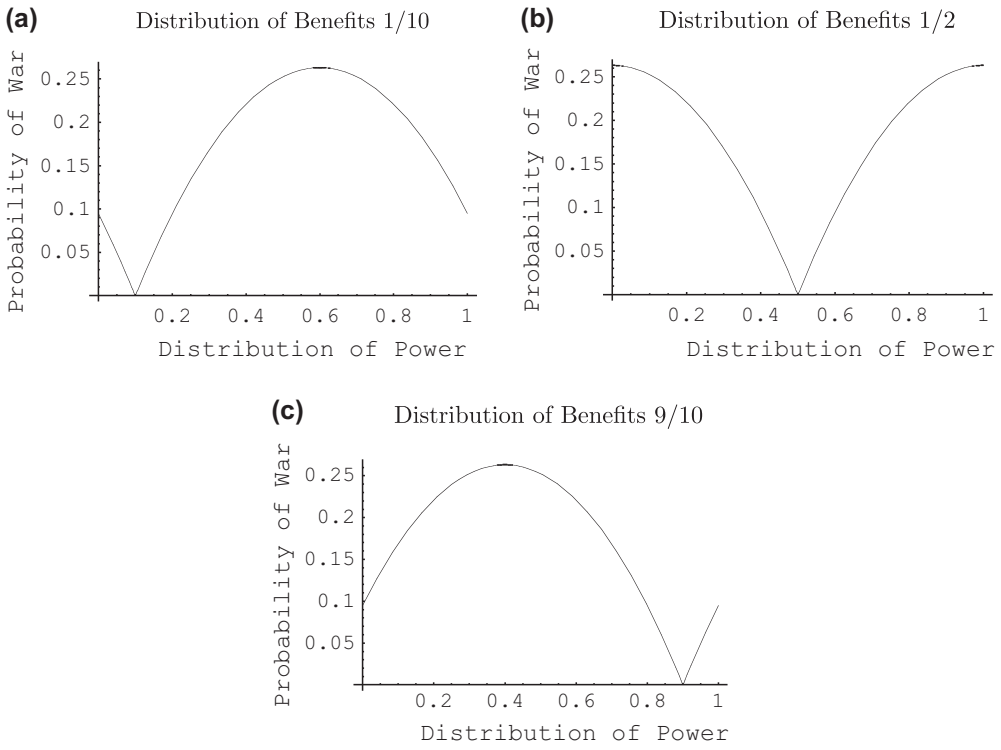
A proxy's value ultimately depends on its usefulness in other applications. In this section, we demonstrate the DOE score's usefulness in two ways: first, through a

[9]The difference in log loss is statistically significant (paired $t = -3.1$, $p = .008$).

[10]We also construct undirected predictions; see the SI for details.

[11]About 19% 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 SI for details.

FIGURE 3 Reed and Colleagues' (2008) Graphical Summary of their Main Hypotheses



detailed replication of a well-known test of the bargaining model; 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 which measure(s) to include.

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 the probability that one would prevail over the other in a conflict, p, and the distribution of benefits between the states, q. For example, q may capture where a border is drawn between two neighbors. 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, 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. (2008) test their theory by incorporating proxies of $|q - p|$ and $(q - p)^2$ (both lagged 1 year) into a model of dispute onset. Their measure of q, the distribution of benefits, is based on United Nations roll-call votes. Their measure of p, the dyadic balance of power, uses material capabilities; it is a normalization 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.

[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.

TABLE 5 Replication of Table 1, Model 1 of Reed et al. (2008, 1213)

	Reed et al. (2008)		DOE Replication	
Variable	**Coefficient**	**SE**	**Coefficient**	**SE**
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	0.489	0.147
$(q - p)^2_{t-1}$	−0.617	0.196	0.451	0.169
Intercept	−1.580	0.026	−1.573	0.025
N	427,904		427,904	
AIC	12030.916		11803.817	
PRL	0.242		0.257	

Note: The unit of analysis is the dyad-year, and the dependent variable is the onset of a militarized interstate dispute.

Using DOE scores, 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 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's fitting better ($Z = 7.80$, $p < 0.001$). The DOE model is also superior according to the Akaike information criterion 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 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 and colleagues' (2008). 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—that is, 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.

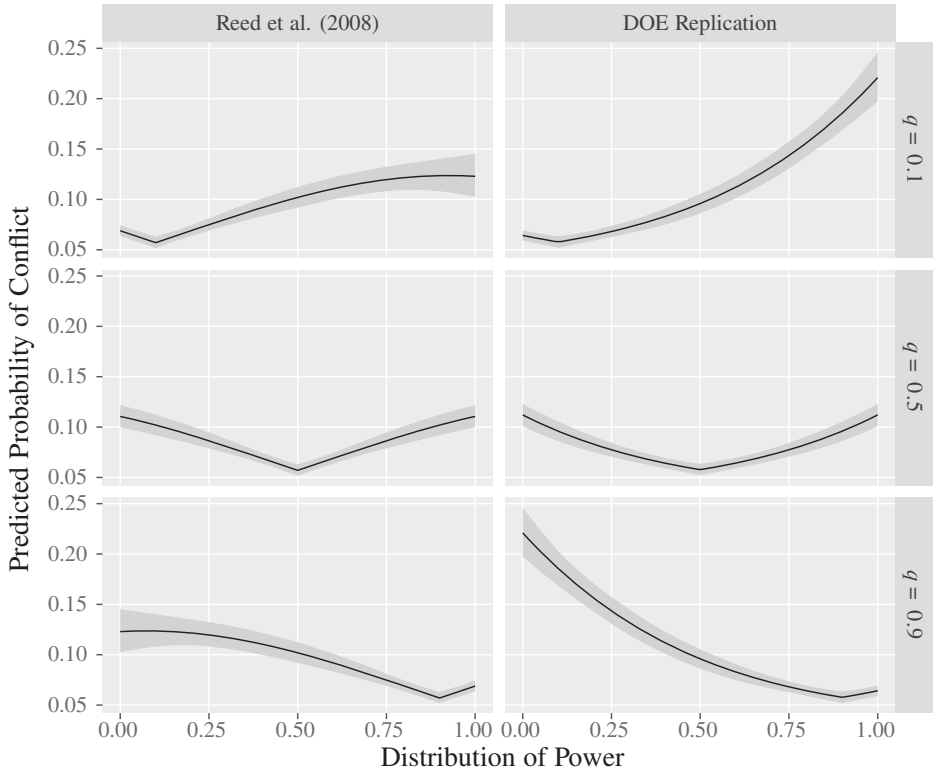
The findings of the reanalysis, which show that parity is never the riskiest distribution of power, run counter to many studies in the dyadic tradition. Indeed, in the first page of the classic dyadic study, Bremer (1992, 309) notes that "a good deal of theoretical speculation and some empirical evidence 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 for such claims, and our results suggest

[14]We reconstruct Reed and colleagues' (2008) measure of p with the latest National Material Capabilities data, so our sample size is slightly larger than in the original article. 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.

[15]See Powell (1999, chap. 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.

FIGURE 4 Replication of Figure 4 of Reed et al. (2008, 1213)



Note: The replication estimates predicted probabilities of conflict while holding democracy and distance at their minimal values. The 95% confidence intervals were obtained via a parametric bootstrap.

that the DOE score might 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 when 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 (1996) original analysis and Reed and colleagues' (2008) 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.

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—that is, 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 is a key concern 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).

TABLE 6 Summary of Results from the Replication Analysis

			Main Hyp.		Power Hyp.	
Replication	***N***	**Vuong**	p_{CINC}	p_{DOE}	p_{CINC}	p_{DOE}
Bennett (2006)	1,065,755	−13.57	√	√	√	√
Weeks (2012)	766,272	4.33	√	√	√	√
Jung (2014)	742,414	1.62	√		√	√
Park and Colaresi (2014)	379,821	1.66	√	√		√
Sobek, Abouharb, and Ingram (2006)	183,227	3.45	√	√	√	√
Gartzke (2007)	171,509	4.13	√	√	√	√
Salehyan (2008b)	86,497	1.21	√	√	√	√
Fuhrmann and Sechser (2014)	85,306	1.41	√	√		√
Arena and Palmer (2009)	54,403	3.18	√		√	√
Owsiak (2012)	15,806	2.31	√	√	√	√
Zawahri and Mitchell (2011)	12,186	0.76	√	√	√	√
Salehyan (2008a)	10,197	1.49	√	√		√
Fordham (2008)	7,788	−2.18	√		√	√
Dreyer (2010)	5,316	2.54	√	√		√
Huth, Croco, and Appel (2012)	3,826	−1.04	√		√	
Uzonyi, Souva, and Golder (2012)	1,667	1.49	√	√		√
Weeks (2008)	1,582	1.19	√	√		√
Morrow (2007)	864	−2.48	√	√	√	√

Note: 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 $p < .05$ for the main substantive hypothesis test under each model; the final two report whether $p < .05$ for a test of the null hypothesis that all power variables have a coefficient of zero.

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 a set of models to replicate 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.[17] 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.[18] 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 article, or at least a representative one.[19] We then re-estimate the original model and a replication in which 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.[20]

To see when the alternative measure results in a different substantive conclusion, we identify the main hypothesis of each study and perform the corresponding null hypothesis test on both the original model and the DOE score replication. Additionally, to test for an effect of relative power, in the original models we test the null

[17]See note 3 above.

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

[19]For details, see the SI.

[20]We employ the standard Bayesian information criterion (Schwarz 1978) correction to the Vuong test statistic. The SI reports additional evaluations; the results are substantively identical.

hypothesis that all CINC-derived terms have a coefficient of zero, and in the replication models we do the same with the DOE terms.

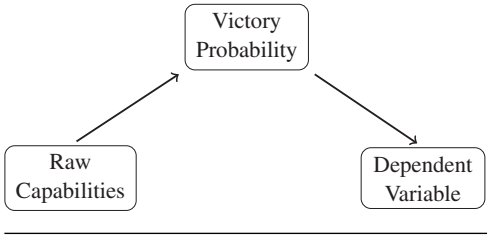
Table 6 summarizes the results of the replication analysis. The results further support our contention that DOE scores are superior to the capability ratio as a proxy for relative power. In a majority of the conflict studies we replicate, we explain more of the variation in the dependent variable when 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 in which 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 $\chi^2 = 6.5$, df $= 8$, p $= .59$).[21] 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 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 $\chi^2 = 2.78$, df $= 2$, p $= .25$). 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 all but 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.

[21]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 SI for details.

FIGURE 5 Raw Capabilities Only Affect the Outcome of Interest through the Probability of Victory



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.

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 by Reed et al. (2008).

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 6 Raw Capabilities Affect the Outcome of Interest Both Directly and through Expectations

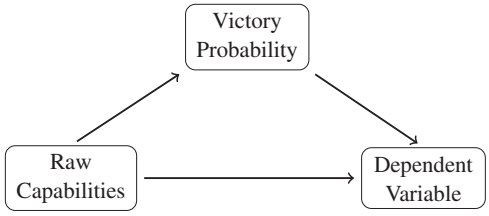
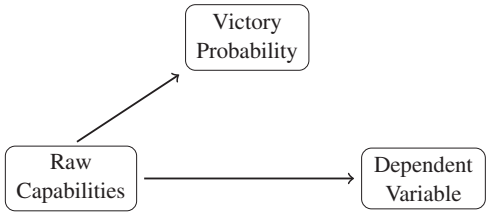


FIGURE 7 Raw Capabilities Directly Affect the Outcome of Interest, But Expectations Do Not



The last possibility to consider is that expectations do not directly affect the outcome of interest, as illustrated in Figure 7. 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 they should not be included as an independent variable when outcome is the dependent variable.

When there is no specific theory about how material capabilities affect the outcome of interest, 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.

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 the 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 when new data are 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 article has developed one measure that improves on the previous state of the art along a number of dimensions.

References

Arena, Phil. 2012. "Measuring Military Capabilities." Blog post. 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): 955–75.

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): 313–38.

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): 309–41.

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. New York: Free Press, 233–57.

Carter, David B., and Curtis S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18(3): 271–92.

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

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2): 355–70.

Dreyer, David R. 2010. "Issue Conflict Accumulation and the Dynamics of Strategic Rivalry." *International Studies Quarterly* 54(3): 779–95.

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): 3133–81.

Fordham, Benjamin O. 2008. "Power or Plenty? Economic Interests, Security Concerns, and American Intervention." *International Studies Quarterly* 52(4): 737–58.

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): 919–35.

Gartzke, Erik. 2007. "The Capitalist Peace." *American Journal of Political Science* 51(1): 166–91.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. New York: Springer.

Hegre, Håvard. 2008. "Gravitating toward War Preponderance May Pacify, But Power Kills." *Journal of Conflict Resolution* 52(4): 566–89.

Hill, Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3): 661–87.

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): 561–81.

Howard, Michael. 1976. *War in European History*. Oxford: 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): 201–17.

Jung, Sung Chul. 2014. "Foreign Targets and Diversionary Conflict." *International Studies Quarterly* 58(3): 566–78.

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: University of Michigan Press, 3–34.

Linzer, Drew, and Jeffrey K. Staton. 2014. "A Measurement Model for Synthesizing Multiple Comparative Indicators: The Case of Judicial Independence." Working paper. 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): 103–20.

Morrow, James D. 2007. "When Do States Follow the Laws of War?" *American Political Science Review* 101(3): 559–72.

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 data set, 2002-2010: Procedures, Coding Rules and Description." *Conflict Management and Peace Science* 32(2): 222–42.

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): 118–25.

Powell, Robert. 1996. "Stability and the Distribution of Power." *World Politics* 48(2): 239–67.

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): 1203–16.

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

Salehyan, Idean. 2008b. "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): 461–64.

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, 19–48.

Slantchev, Branislav L. 2003. "The Principle of Convergence in Wartime Negotiations." *American Political Science Review* 97(4): 621–32.

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): 519–29.

Stam, Allan C. 1996. *Win, Lose, or Draw: Domestic Politics and the Crucible of War*. Ann Arbor: 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: 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): 822–29.

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): 1–21.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2): 307–33.

Waltz, Kenneth N. 1979. *Theory of International Politics*. Boston: 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): 326–47.

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): 835–58.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix 1: National Material Capabilities Data
Appendix 2: Militarized Interstate Dispute Data
Appendix 3: Multiple Imputation
Appendix 4: Super Learner Candidate Models
Appendix 5: Undirected DOE Scores
Appendix 6: Replications