Capability Ratios Predict Nothing*

Robert J. Carroll† Brenton Kenkel‡

November 11, 2015

Abstract

Modern approaches to political measurement have generally ignored the importance of out-of-sample predictive performance. This is problematic for two reasons: first, many of the abstractions scholars attempt to proxy for are themselves expectations; and second, the resulting measures are prone to overfitting. We advocate a data-driven approach to measurement focused on out-of-sample prediction. We demonstrate the effectiveness of the approach as it applies to proxying the expected outcome of militarized interstate disputes. The standard proxy for expected dispute outcomes, the ratio of material capability indices, has almost no predictive power. We use ensemble learning to construct a new measure from the same underlying covariates—the Dispute Outcome Expectations score, or DOE—whose predictive power far exceeds that of the standard measure. In replications of 18 empirical studies of international relations, we find that replacing standard capability measures with DOE scores usually improves both in-sample and out-of-sample goodness of fit.

*We thank Scott Bennett, Brett Benson, Inken von Borzyskowski, Kevin Clarke, Mark Fey, James Honaker, Zach Jones, Karen Jusko, Holger Kern, David Lewis, Adeline Lo, Matt Pietryka, Marc Ratkovic, Jim Ray, Mark Souva, and Hye Young You for helpful discussions and advice. Bryan Rooney provided excellent research assistance. We also thank the authors listed in Table 4 for making their replication data publicly available. The current version of the Dispute Outcome Expectations data can be downloaded from https://dataverse.harvard.edu/dataverse/doe-scores. Replication code and a version history of the project are available at https://github.com/brentonk/crpn.

†Assistant Professor, Department of Political Science, Florida State University. Email: RobCarrollFSU@gmail.com.

‡Assistant Professor, Department of Political Science, Vanderbilt University. Email: brenton.kenkel@vanderbilt.edu.

Of all the challenges in political science, perhaps none is more difficult and rewarding than measuring theoretical quantities. Sometimes the most important concepts are the most elusive to measure. Many analyses tackle the measurement problem by using (or producing) the simplest measure possible, often in the form of summated rating scales.[1] The practice persists even though these simple measures often introduce *ad hoc* assumptions (such as those regarding the weights to attribute to each item in the rating scale), and authors are often apologetic for their use. Over the years, methodologists have mitigated old frustrations by developing better models for measuring a variety of quantities, from ideal points (Clinton, Jackman and Rivers 2004) to judicial independence (Linzer and Staton 2014) to democracy (Jackman and Treier 2008). At the same time, scholars have amassed an impressive amount of data, particularly historical data. Ideal points now use roll calls back to the American Constitutional Convention (Heckelman and Dougherty 2013); conflict scholars can access industrial output figures for each state dating back to the Napoleonic Wars (Singer, Bremer and Stuckey 1972). Improvements in computing power, coupled with scholarly ingenuity, ensure that this progress will continue for the foreseeable future.

We should feel sanguine given these advances, but we should also pause to consider the nature of the measurement enterprise. Our new data sets enable us to ask and answer meaningful measurement questions, and this often means doing more than taking unweighted averages of our new variables. Put differently, when imposed on new and interesting data, crude measures seem especially crude—we cannot even complain that they are underfit, since they are not fit to data at all. At the other extreme, measurement models with an abundance of parameters run the risk of overfitting: the attribution of systematic importance to random error. Likewise, as our data sets grow, so too do we run the risk of attributing too much reliability (sociologically speaking) to our potentially overfit results. Yet, to our knowledge, none of the recent advances in political measurement have taken out-of-sample performance into account; rather, attention is paid to developing and interpreting measures that best reflect extant data.[2] We aim to provide an approach to measurement that strikes a balance between the severely underfit approaches common to applied work and the potentially overfit approaches developed by specialized technicians. Doing so requires us to pay close attention to prediction from the outset.

[1] For a brief introduction, see Spector (2006).

[2] In contrast, structural modelers have paid increasing attention to overfitting problems (Pitt and Myung 2002; Preacher 2006), though most instruction retains its focus on fit. We should note that some have examined how *unsupervised* learning relates to prediction; see Tibshirani and Walther (2005).

Though any measure suffers when a data set's unique peculiarities are assigned too much explanatory import, theoretical expectations provide unique challenges. To animate the situation, imagine a real-world leader that must decide whether to start a war against another state. Perhaps inspired by Santayana's observation that "those who cannot remember the past are condemned to repeat it," the leader orders her statisticians to obtain data on the outcomes of previous conflicts and the material capabilities of their combatants. The statisticians, of course, could use the data in a variety of ways to produce a prediction, but the leader would care only that the prediction was the one that did the best job of predicting. More to the point, the leader would care less that estimates of the relevant parameters fit the historical data as well as possible (as would be the case if the statisticians ran traditional logit or probit models alone) and would care more that the prediction was of high quality. Indeed, to borrow another aphorism, we can reimagine the oft-lamented sin of "fighting the last war" (e.g. Hart 1972) as overfitting such historical models with excess weight placed on recent observations. Yet, when we write down formal models of choice under uncertainty featuring actors like this leader, we operate on the assumption that the expectation in question is, by definition, the predictor of the relevant outcome—and, as such, is neither under- nor overfit to some other relevant data at the hypothetical decision-maker's disposal.

Just like the leader, empirically-minded scholars want the estimate of war outcomes that predicts best rather than the one that fights the last war. Consider the statistical model of militarized interstate dispute onset analyzed by Leeds (2003). Though she is interested in the effect of outside alliances on dispute onset, Leeds argues that she "must embed [alliance] variables in an empirical model that predicts a base probability of dispute initiation" (433). One contributor to such a baseline model is a variable that "compares the power of the potential challenger to the power of the potential target," which is justified "because stronger states are more likely to *expect military success*" (434, emphasis added). Unsurprisingly, then, Leeds constructs the ratio of the capabilities of one state to the sum of the capabilities of the dyad, which ranges from 0 to 1—as a probability does, highlighting the measure's theoretical roots as an expectation. The capability measures, Composite Indices of National Capabilities (Singer, Bremer and Stuckey 1972), are themselves transformations of summated rating scales that were constructed *a priori*, without data-driven choices of weights or transformations. But if we were to construct a traditional measurement model for the CINC scores' underlying variables, we would run the risk of overfitting.

We aim to do right by both the leader and the scholar. In this article, we argue that proxies should be constructed to predict well and that functional forms

should be assessed on that criterion. We advocate a data-driven approach focused on out-of-sample prediction: a proxy for the expectation of some political outcome ought to be a good predictor of that outcome. When selecting from the variety of potential models to construct a proxy variable, the data used to *assess* the model should not be same as the data used to *fit* it. Techniques that accomplish this division of labor through sample-splitting, such as cross-validation (Efron and Gong 1983), ought to be more widely used in measurement construction. Our arguments in favor of predictive power mirror those 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 measures rather than comparing them—in particular, we examine how to create proxy variables with the greatest ability to predict.

We apply our approach to the measurement of political power, which arises in all areas of political science but is especially important in the study of international conflict. The bargaining model of war (Fearon 1995)—long the workhorse model in modern IR theory—operationalizes power into expected dispute outcomes, most often represented by the probability that one state defeats another in battle, denoted p. Given the bargaining model's importance, it comes as no surprise that empirical scholars have tried to proxy for p in their analyses. Most scholars have followed Leeds' lead and used ratios of CINC scores, more commonly called capability ratios.[3] Capability ratios are inappropriate as a proxy for expected dispute outcomes for a variety of reasons, including problems with the CINC function itself, *ad hoc* parameterizations, and issues of functional form. What is more, when evaluated on predictive performance, we find that *capability ratios predict nothing*: they fail to predict any outcome other than the modal category (that the dispute ends in a stalemate) and barely improve out-of-sample predictive performance (0.8%) over a null model. Conversely, our method yields predictions with much more variation and improves out-of-sample predictive performance by 20%. It is notable that this is the case even though we use the same component variables from which the CINC score is constructed.

Our result is also of major substantive interest: a scholar armed only with capability ratios and a logit routine would conclude that material capabilities have little effect on dispute outcomes, whereas our results suggest that material capabilities matter in a very real way. That effect can only manifest itself,

[3] In a search of some of the top journals for empirical work in international relations, we found at least 94 articles between 2005 and 2014 using CINC ratios or another function of CINC scores in a dyadic analysis. The journals examined were *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *International Organization*, and *International Studies Quarterly*.

however, when the analyst uses tools that allow subtle and nuanced relationships among the predictors and the response to shine through—that is, through the use of flexible algorithms. When deciding whether to roll the iron dice, decision-makers make calculations over the evidence in a way that we as analysts could never hope to understand through statistical machinery alone. But at the very least, our analysis suggests that statistical flexibility can produce results that satisfy both the historical record and our scholarly priors.

We use our method to construct a new measure: the DOE (Dispute Outcome Expectations) score. DOE scores retain the simple, probabilistic flavor advanced in Leeds' justification for inclusion in her model. For every dyad-year (or directed-dyad-year) covered by the Correlates of War data, we provide the probability that each state would win a hypothetical dispute as well as the probability of stalemate. Unlike other contrivances based on CINC ratios, our scores make intuitive sense when interpreted in plain language.

In advocating for the DOE score, we are not asking applied international relations scholars to give anything up. We replicated 18 recent empirical studies that utilized the capability ratio and then replaced it with the DOE score. In 14 of the 18 replications, the DOE score improved out-of-sample goodness of fit, and it improved in-sample fit 15 times. All told, the DOE score is preferable than the capability ratio for four reasons: it more directly reflects theories of international relations; it avoids the underfitting of *ad hoc* measures while avoiding the overfitting of traditional measurement models; it has a natural, probabilistic interpretation; and it usually performs better in the kinds of analyses most applied scholars care about.

The paper proceeds in six sections. In the first two, we argue for the importance of predictive power in constructing proxies and assess the functional problems associated with capability ratios. Section 3 describes the data and methods we use to construct a new proxy for expected dispute outcomes. In Section 4, we discuss the advantages and disadvantages of our new measure, the DOE score. Section 5 provides the results of our replications. The final section addresses next steps and concludes.

1 Predictive Power and Proxy Variables

We are often interested in questions that link observed data to some unobserved quantity. This latter quantity may be unobserved because it is difficult (or impossible) to measure directly (like wealth) or because it is an abstraction (like the ideal point of a voter in a spatial model). In either case, the applied analyst faces a choice between omitting some potentially important variable and includ-

ing some proxy variable in its stead (Stahlecker and Trenkler 1993). There is no best choice: some theoretical econometricians (e.g. McCallum 1972) argue for the inclusion of all proxies (including crude ones), while others (e.g. Maddala 1977) support only the use of reliable proxies. Even those in the former camp, however, admit that reliable proxies perform better than unreliable ones.

Healthy disciplines use good measures for central concepts (Kuhn 1977), and so social science progresses, in part, by developing better ways to construct proxy variables.[4] Much recent progress is due to the development of measurement models.[5] Jacoby (2014, 2) observes:

> "All of us are comfortable with the notion of statistical *models* that provide representations of structural relationships between variables. But, modern social science also regards measurement as a model that pertains to each of the individual variables. Careful attention and rigorous approaches are just as important for the latter type of models, as they are for the former."

Moreover, when the unobserved quantity is an abstraction, appropriate measurement models allow the analyst to perform direct tests that follow from the same set of assumptions as those used in the original, theoretical model. As Clinton, Jackman and Rivers (2004, 355) put it in the context of testing legislative behavior, "it is inappropriate to use ideal points estimated under one set of assumptions...to test a different behavioral model...." Shor and McCarty (2011, 530) note that the close relationship between statistical and theoretical models of legislative behavior (and their requisite assumptions) "has contributed to a much tighter link between theory and empirics in these subfields of political science."

While better models (and better ways to estimate their parameters) have improved our measures of a variety of important quantities, it remains problematic that modern political measurement has ignored the importance of predictive power in producing proxies. This is odd, as seminal contributions to the literature utilize classification—a criterion often used in machine learning, where the focus is usually on prediction—as a way to prove a new measure's

[4] Here we focus on the importance of models in producing measures; equal weight should be assigned to advances in the estimation of these models' relevant parameters, most notably to advances in Bayesian estimation (Jackman 2001; Martin and Quinn 2002; Clinton, Jackman and Rivers 2004; Bafumi et al. 2005).

[5] Of course, the use of theory in the act of measurement is nothing new. Economics retains its longstanding commitment to structural estimation whereby theoretical models are used to uncover relevant quantities. For current applications to the structural estimation of dynamic discrete-choice games (for example), see Su and Judd (2012) and Egesdal, Lai and Su (2013).

superiority over extant ones. For example, in the classic paper on ideal point estimation in American legislatures, Poole and Rosenthal (1985, Table 3) report that a simple classification approach based on their NOMINATE scores correctly predicts over 80% of legislative votes in most years in their data. Though their procedure estimates ideal points via the method of maximum likelihood rather than via a classification criterion, it remains that this analysis lies prone to textbook overfitting problems. In their paper, Poole and Rosenthal estimate ideal points within a single Congressional session, and their classification test then uses those ideal points to assess voting within the same Congressional session. While many of the correct classifications reflect the spatial model's explanatory virtues, others may arise due to overfitting to the data within that Congressional session.

It is for these reasons that we advocate a data-driven approach based on predictive performance. While traditional statistical approaches to measurement minimize error or maximize likelihood within the entire data set, we instead aim to optimize out-of-sample prediction. We do so through cross-validation, a statistical learning procedure that allows us to simulate out-of-sample prediction without withholding any data in the model fitting process. We fit a variety of models to the data, including highly flexible machine learning algorithms, and assess each of their predictive performance through cross-validation. We then make our final predictions by taking a weighted average of these many individual models. This approach explicitly addresses the problems enumerated above: it avoids the overfitting problems associated with traditional measurement techniques and the model selection problems associated with attempts to bring new data to bear for predictive purposes. We believe that the costs of our approach—additional computation and difficulty in interpreting the results—pale in comparison to these benefits.

2 The Capability Ratio and Its Discontents

Thanks in part to the popularity of formal models of choice under uncertainty, many unobserved quantities like those discussed above take the form of probabilities. Our application—expectations about war outcomes as parameterized by some probability $p \in [0, 1]$—is no different. We want to create a proxy for the chance that Country A would prevail in a dispute against Country B, given their observable characteristics, x_A and x_B. Since a measure of a probability must lie within the unit interval, a natural way to proceed is to propose an

indexing function g, where $g(x) \geq 0$, and then take the ratio of indices,

$$f(x_A, x_B) = \frac{g(x_A)}{g(x_A) + g(x_B)}. \tag{1}$$

The quality of such a measure depends on both the selected characteristics and on the appropriateness of the indexing function g. This latter responsibility plays a large role in the development of good measures and is our primary area of focus.

Though simple, this enhanced ratio-based approach is remarkably powerful and finds use in a diverse array of applications. A classic success comes from the study of baseball outcomes, where the Pythagorean prediction (James 1983; Miller 2007) of a team's winning percentage is defined as

$$f\left(\text{Runs Scored}, \text{Runs Allowed}; \alpha\right) = \frac{\text{Runs Scored}^{\alpha}}{\text{Runs Scored}^{\alpha} + \text{Runs Allowed}^{\alpha}},$$

where $\alpha \geq 0$ adjusts x's shape. Here the quest for the best-fitting g amounts to estimating α; James (1983) originally proposed $\alpha = 2$ *ad hoc*, and later analysts found that $\alpha = 1.83$ fit the data best. Though the analyses that produced this estimate suffer from the overfitting problems discussed above, the Pythagorean predictor still performs quite well when imposed upon out of sample data.

When proxying for expected dispute outcomes, empirical conflict scholars normally use transformations of data on states' material capabilities. We now relate the typical transformation to our discussion of ratio-based measures above. We begin by introducing some helpful notation: call the set of states $\mathcal{I} = \{1, \ldots, I\}$; the set of variables $\mathcal{J} = \{1, \ldots, J\}$; and the set of years $\mathcal{T} = \{1, \ldots, T\}$. Denote state i's value for variable j in time t as M_{ijt}. The set of all data is M, and all data in year t is M_t. Define state i's share of variable j in year t as

$$S_{ijt}(M_t) = \frac{M_{ijt}}{\sum_{\mathcal{I}} M_{ijt}}.$$

We now introduce the *CINC function*.[6] State i's CINC score in year t is its average share across all variables:

$$\text{CINC}_i(M_t) = \frac{\sum_{\mathcal{J}} S_{ijt}}{|\mathcal{J}|}.$$

[6]Here CINC stands for "Composite Index of National Capability" as given in the Correlates of War National Material Capabilities data (Singer, Bremer and Stuckey 1972).

State i's CINC score therefore falls in $[0,1]$. The CINC score is the "most commonly used measure" of power in empirical conflict studies (Kadera and Sorokin 2004, 212).

Following the discussion in the previous section, the most intuitive CINC-based proxy for p is a naïve *capability ratio*:

$$f_{\text{CR}}(M_t) = \frac{\text{CINC}_A(M_t)}{\text{CINC}_A(M_t) + \text{CINC}_B(M_t)}.$$

Our approach, then, makes explicit the fact that CINC is simply a candidate g function imposed upon annual material capability data M_t.

Problems emerge immediately. Two of these pertain to the CINC function itself. First, as has been documented, the CINC function is sensitive to changes in state membership over time (Organski and Kugler 1980; Gleditsch and Ward 1999; Kadera and Sorokin 2004). Second, the CINC function's equal weighting of all indicators is entirely *ad hoc*. For example, the CINC function assigns the same importance to military spending as it does to personal energy consumption. Whether this is an appropriate assignment is an empirical question that goes unanswered. Even on the tenuous assumption that CINC is a good data reduction technique on M_t, it is not clear whether it serves as a good g function. Prior to entering f_{CR}, should the CINC scores be exponentiated given some parameter on returns to scale, or perhaps instead logged? Finally, even given that CINC is a useful index, we do not know whether a ratio-based approach is appropriate at all. In other words, taking the capability ratio at its word requires making a host of assumptions that may not hold well enough to make it useful in applications.[7]

Yet it is widely used. Capability ratios (or similar manipulations of CINC scores) feature prominently in many recent empirical studies in international relations. As we might expect given the importance of the bargaining model, many of these (e.g. Gartzke 2007; Salehyan 2008*a*) use capability ratios in regressions predicting the onset of a militarized interstate dispute. Still others focus on particular features of a militarized interstate dispute, such as the nature of its termination (Beardsley 2008) or whether its combatants complied

[7] It is worth noting that some applications follow Bremer (1992) in using the explicit CINC ratio:

$$f_{\text{Bremer}}(M_t) = \frac{\max\{\text{CINC}_A(M_t), \text{CINC}_B(M_t)\}}{\min\{\text{CINC}_A(M_t), \text{CINC}_B(M_t)\}}.$$

Bremer's approach does not fall in $[0,1]$, though it could be transformed through a logit or probit CDF. However, it is a monotonic transformation of the aforementioned capability ratio, and it suffers from similar problems anyway.

with laws of war (Morrow 2007). Still other studies focus on other phenomena not directly related to disputes, such as the onset of sanctions (Whang, McLean and Kuberski 2013), issue agreements (Mitchell and Hensel 2007), or nuclear assistance provisions (Kroenig 2009). A more exhaustive survey of the use of capability ratios is beyond the scope of this paper, but suffice it to say that it is the go-to measure of relative power in international relations.

One might politely defend the capability ratio by noting that it asks the CINC function to perform a job it was not designed for. We would agree. It is worth noting, however, that early proponents of the CINC function (Singer, Bremer and Stuckey 1972, 24) sought to understand how "uncertainty links[s] up with capability patterns on the one hand and with war or peace on the other." Writing over two decades before the classic introduction to the bargaining model of war (Fearon 1995), these authors lacked the abstract target—p—that we currently have, but their enterprise was largely similar. Though their focus on systemic, rather than dyadic, patterns reflects the dominant flavor of realism at the time, they still wanted to know how preponderance of power related to the decision to fight. So, while it remains true that the capability ratio is not meant to directly relate to p in a bargaining model, it *was* meant to tell us something about how uncertainty relates to war.

Remedying these problems by fitting a model—perhaps estimating an exponent on the CINC scores in f_{CR}, or weights for the CINC indicators—would be a laudable first step, but it would still suffer from functional form dependencies. We can sidestep functional form selection by fitting flexible models, but in doing so we must be careful to separate the signal from the noise and keep from overfitting. In the next section, we outline our method for estimating p that suffers neither from overfitting nor from pathologies in the CINC-utilizing, ratio-based approach.

3 Building a Better Proxy for Expected Dispute Outcomes

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*b*). So, to build our new measure, we augment traditional approaches with methods from machine learning.

3.1 Data

To evaluate the predictive performance of the capability ratio and then to build an alternative measure, we use data on the outcomes of international disputes. 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.[8] Every dispute outcome is either A Wins, B Wins, or Stalemate, which we denote by $Y_i \in \{A, B, \emptyset\}$, respectively. 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.

	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.[9] 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. The values of these predictors for the i'th dispute are collected in the vector X_i.

3.2 A Metric for Predictive Power

We face two challenges in evaluating a model's predictive power. The first is to define a metric for predictive power—one that is appropriate to the task at hand and reasonably interpretable. The second is to measure each model's ability to

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

[9] There are missing observations in the National Material Capabilities data. Consequently, 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.

predict *out of sample*. Our main purpose, which is to measure the chances of victory for each side in a hypothetical interstate dispute, is inherently an out-of-sample prediction task. We do not want a model that overfits the sample data at the expense of its predictive power when brought to new data.

As fortune plays a role in every military engagement, it would be 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 $\hat{f}$ is there to emphasize that the form of the function has been learned from the data, whether by estimating regression coefficients or by a more flexible predictive algorithm. The log loss of the model $\hat{f}$ on the data (X, Y) is[10]

$$\ell(\hat{f}, X, Y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\{A,B,\emptyset\}} \mathbf{1}\{Y_i = t\}\log \hat{f}_t(X_i). \tag{2}$$

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. Our small sample size of $N = 1{,}740$ makes this tricky. If we had a surplus of observations, we could use some suitably large number to fit our models and hold out the remainder to assess the models' predictive power. But with as little data as we have, splitting the sample is ill-advised: we cannot hold out enough observations to estimate the generalization error precisely without harming the precision of the model itself. So, to measure out-of-sample predictive power without losing data, we turn to K-fold cross-validation (Hastie, Tibshirani and Friedman 2009, 241–249). We randomly assign each dispute observation to a "fold" $k \in \{1, \ldots, K\}$, where we follow standard practice by setting $K = 10$.[11] For each k, we split the data into a "test" sample containing fold k and a "training" sample containing the remainder of the data. We fit a model only on the

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

[11] Standard practice here stands on firm ground; Molinaro, Simon and Pfeiffer (2005) find that 10-fold cross-validation performs quite similarly to leave-one-out cross validation (the "ideal case" for cross-validation) without having to take on massive computational costs. 10-fold cross-validation also performs better than the .632+ bootstrap, split-sample techniques, and Monte Carlo cross-validation, particularly in smaller samples like ours.

training sample and then calculate its predicted probabilities for the data in the test sample.[12] After repeating this K times, we have an out-of-sample prediction for each observation in our data—one that was calculated from a model that did not see the observation in question. We compare these predicted probabilities to the observed outcomes to estimate our models' generalization error. Formally, the cross-validation loss of the model $\hat{f}$ is the average out-of-fold log loss,

$$\text{CVL}(\hat{f}) = \frac{1}{K}\sum_{k=1}^{K} \ell\left(\hat{f}^{(-k)}, X^{(k)}, Y^{(k)}\right), \tag{3}$$

where $(X^{(k)}, Y^{(k)})$ is the data in the k'th fold and $\hat{f}^{(-k)}$ is the model $\hat{f}$ fit to the data excluding the k'th fold.

Because it is measured on the log-likelihood scale, the log loss metric is hard to interpret on its own. To ease the 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

$$\text{PRL}(\hat{f}) = \frac{\text{CVL}(\hat{f}_{\text{null}}) - \text{CVL}(\hat{f})}{\text{CVL}(\hat{f}_{\text{null}})}. \tag{4}$$

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

3.3 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 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

[12] 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. This nested cross-validation is necessary to keep our estimates of generalization error from being too optimistic (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. Because there are no missing values in the CINC scores, these estimates are identical across imputed datasets.

capability ratio is statistically significant but small relative to the cutpoints, indicating a substantively weak relationship. Dividing the cutpoints by the coefficient, we see that we would need a logged capability ratio below −13 or above +7 to predict any outcome other than a stalemate. These bounds lie well outside the observed range of capability ratios in the dispute data, which are bounded below by −9.1 (Palau–Philippines 2000) and above by −0.0004 (Germany–Panama 1940). In other words, the capability ratio 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 data-generating process—the true relationship between material capabilities and dispute outcomes—looks like. So we use tools from machine learning that are designed to predict well without imposing much structure on the data. Ideally, we would select the predictive model that is best for our data, but there are too many algorithms to try them all. To narrow it down, we defer to the machine learning experts on which algorithms are best. We draw our 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,[13] we end up with six predictive algorithms to try: C5.0, support vector machines, k-nearest neighbors, classification and regression trees, random forests, and ensembles of neural nets.[14] 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.

[13] 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.

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

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 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. All told, we have 30 candidate models: four sets of variables for each of our seven algorithms, plus the capability ratio 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, once we have the cross-validation results, is to choose a model—the one we will use 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

$$\begin{aligned} \min_{w_1,\ldots,w_M} \quad & \text{CVL}\left(\sum_{m=1}^{M} w_m \hat{f}_m\right) \\ \text{s.t.} \quad & w_1, \ldots, w_m \geq 0, \\ & w_1 + \ldots + w_m = 1, \end{aligned} \tag{5}$$

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.[15] That said, the super learner does provide the capability ratio with an opportunity to defend itself; should it earn a high weight, then our costly enterprise may not be worth the effort.

To summarize, we fit and cross-validate $M = 30$ candidate models, then combine them into a super learner that we will use to construct a better proxy for expected dispute outcomes. The biggest downside of our approach is that the results are not easily interpretable. Because the super learner entails av-

[15] As usual when selecting tuning parameters via cross-validation, the value of equation (5) 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.

eraging a large set of models—some of which, like random forests, are themselves difficult to interpret—it gives us no simple summary of how each predictor affects dispute outcomes. This is not a problem, given our aims. Certainly, we would not recommend the super learner as a means of testing hypotheses about the determinants of dispute outcomes. However, our goal is not to test a hypothesis—it is to construct the best proxy possible for how a dispute between two countries is likely to end. In this context, it is worth sacrificing interpretability for the sake of predictive power.

3.4 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 27 alternative models has 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.

We can glean a few basic intuitions about the relationship between capabilities and dispute outcomes from the cross-validation results. First, the relationship is too complex to capture in a linear model. The best of our ordered logits has a 10 percent proportional reduction in loss, about half that of the best nonlinear model (the neural net ensemble on capability components with year included). So we cannot fix the problems with the capability ratio just by disaggregating it into its component parts. Second, the capability–outcome relationship changes over time. Accounting for the year of the dispute improves our models' predictive power: in 13 out of 14 cases, the model that includes time predicts better than its time-less counterpart.[16] Unlike the capability ratio, flexible predictive algorithms can capture this variation. Finally, the models trained on the raw capability components tend to outperform those trained on the annual shares of components.[17] This runs contrary to way the CINC score is constructed—it is an average of capability shares—suggesting another reason why the capability ratio is such a poor predictor.

As we expected, the super learner ensemble performs better than any of the candidate models from which it is constructed. The ensemble's proportional

[16] The difference in log loss is statistically significant (paired $t = -3.25$, $p = 0.006$).
[17] The difference in log loss is statistically significant (paired $t = -3.14$, $p = 0.008$).

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	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	0.02	0.01
C5.0	Components	✓	0.51	0.04	0.04
C5.0	Proportions		0.52	0.03	0.02
C5.0	Proportions	✓	0.51	0.05	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	0.10	<0.01
k-Nearest Neighbors	Components		0.47	0.12	<0.01
k-Nearest Neighbors	Components	✓	0.45	0.16	0.02
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	0.19	0.28
CART	Proportions		0.55	−0.03	<0.01
CART	Proportions	✓	0.50	0.06	<0.01
Random Forests	Components		0.49	0.08	0.04
Random Forests	Components	✓	0.48	0.11	0.19
Random Forests	Proportions		0.47	0.12	<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	0.19	0.13
Averaged Neural Nets	Proportions		0.48	0.11	<0.01
Averaged Neural Nets	Proportions	✓	0.44	0.19	0.16
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.

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 12 and 15), the super learner's predictive power is still the 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.

The super learner predicts dispute outcomes much better than the capability ratio does. As we have just shown, the capability ratio only improves by 1 percent on a null model, whereas the super learner gives a 20 percent improvement. For another illustration of the difference in predictive power, see the plots of out-of-fold predicted probabilities—the ones we use in cross-validation—in Figure 1. 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.

What should we make of the fact that, even after all this predictive effort, about 80 percent of the variation in dispute outcomes remains unexplained? One problem is that our sample size is too small to fit flexible models without a loss of precision. For the sake of precise predictions (though certainly not that of humankind), it would be better to have far more dispute observations to use for model building. The bigger issue, however, is that material capabilities do not tell the whole story. No matter how many modeling tricks we pull out, we cannot change the fact that militarized dispute outcomes depend on much more than raw capabilities. Some of these other predictors, like the distance between the disputants, are easily measured; others, like the cleverness of their leaders, are not. In principle, it is possible to incorporate additional variables (at least those that can be quantified) into the super learner to attain even better outcome predictions. That goes beyond the scope of this paper, which only aims to show how we can best use the data underlying the capability ratio.

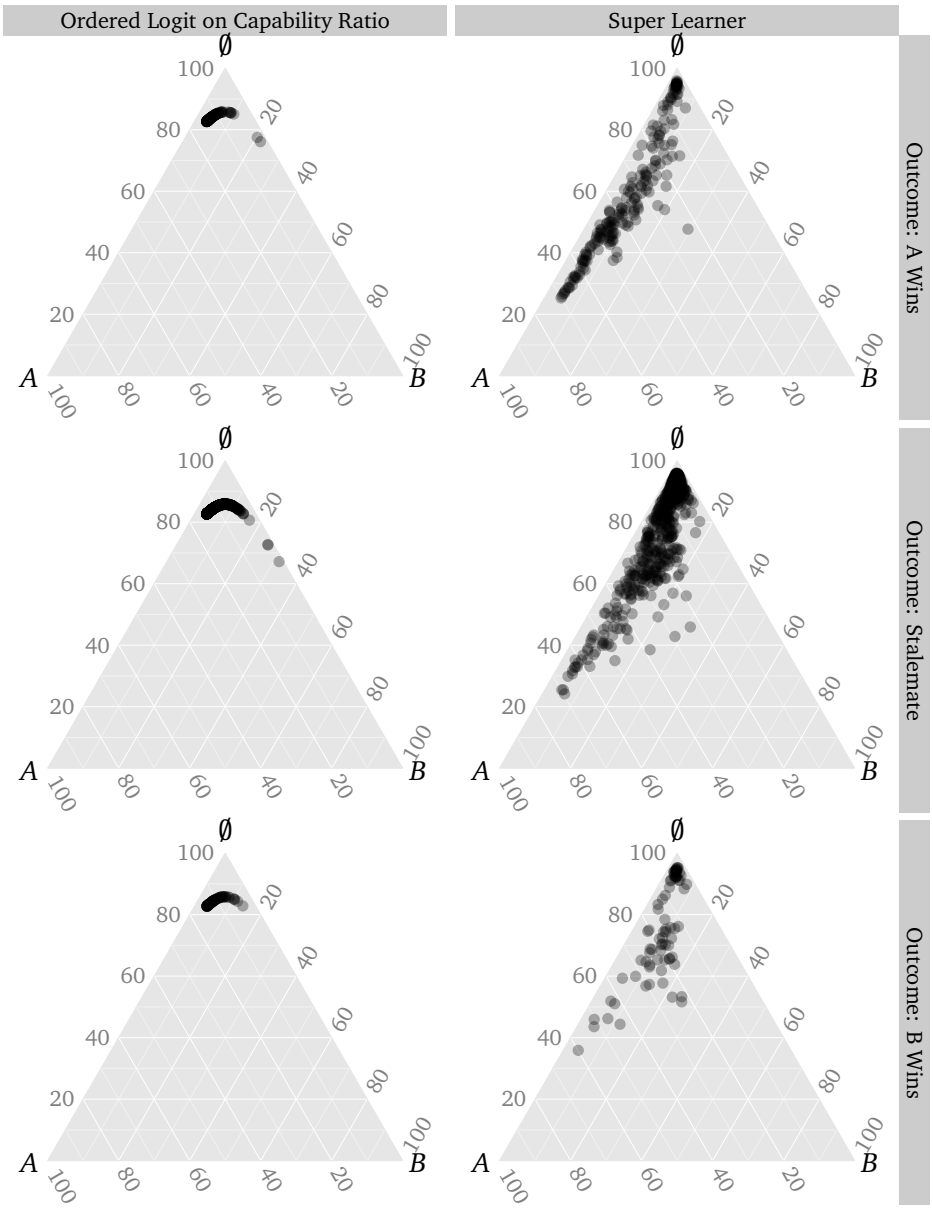

Figure 1. Ternary plots of out-of-fold predicted probabilities according to the capability ratio model and the super learner. Each predicted probability is calculated by fitting the model to 9/10 of the data, not including the observation in question—an approach that simulates true out-of-sample prediction.

4 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.[18]

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 (50 percent) and as a target (10 percent) to yield 30 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.

The DOE measures have two advantages over the capability ratio as a proxy for expected dispute outcomes. First, they are direct measures of the quantity of primary interest to scholars of conflict: the probability that each state would win in a hypothetical dispute. Although the capability ratio is a proportion, it cannot be interpreted as the probability of victory. The ease of interpretation is particularly important for scholars who wish to control for expected dispute outcomes in a regression model. The coefficient on a DOE score can be interpreted directly as the marginal effect of a state's chance of victory; the coefficient on the capability ratio cannot. Second, as we have already seen, within the set of state pairs where disputes occur, the DOE measures are much better predictors of the outcome than the capability ratio is. In short, they are superior proxies, and therefore are the appropriate choice for scholars who need an accurate measure of expected dispute outcomes. The canonical correlation between the

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

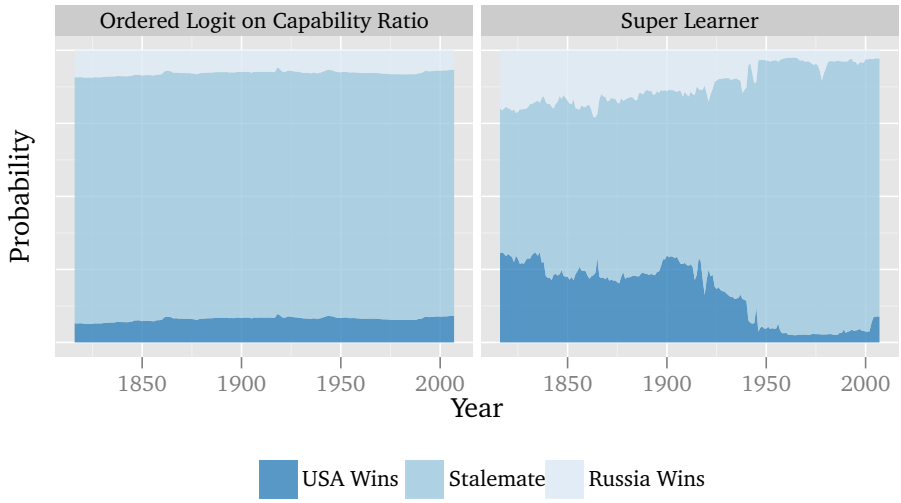

Figure 2. Comparison of predicted outcome probabilities over time from the capability ratio and the super learner (DOE scores) for the United States versus Russia. These plots use the undirected scores.

DOE scores and the capability ratio is 0.44 (for both the directed and undirected DOE scores), so the measures are related but distinct.

The DOE scores have one drawback worth mentioning: they 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. Another way to think about it is that the DOE score measures expectations of dispute outcomes, and there is no reason to think these expectations themselves have an independent effect on the outcomes. So when we test causal hypotheses about dispute outcomes, we should control for raw capabilities, not expectations. But when we are modeling dependent variables that might be affected by expectations, such as the onset of a crisis or a state's decision to join an ongoing conflict, we should use the best available proxy for those expectations—namely, the DOE scores.

To illustrate the contrast between the capability ratio and DOE scores for

forecasting hypothetical dispute outcomes, Figure 2 plot the two models' predictions over time for the United States versus Russia. According to the capability ratio model, there was essentially no change between 1816 and 2007 in the likelihood of either side winning a dispute. We need not dwell on the implausibility of this prediction. Conversely, the DOE scores stack up with our intuitions relatively well: the predicted chance of a stalemate balloons during the Cold War, but the chance of victory by the United States picks up afterward.

In light of the DOE scores' superior predictive performance in the Militarized Interstate Disputes data, we are inclined to believe they dominate the capability ratio as a proxy for expected dispute outcomes. Next, we test this conjecture by seeing if replacing the capability ratio with DOE scores in empirical models of international conflict improves their in-sample fit and out-of-sample predictive power.

5 Using the New Measure

So far, we have focused on showing that DOE scores predict militarized dispute outcomes better than the capability ratio does. However, international relations scholars often control for the capability ratio as a proxy for expected outcomes when modeling dependent variables besides who wins, such as the onset or escalation of a dispute. We have shown that the DOE score is a better proxy than the capability ratio, but it does not follow immediately that it is a better control variable in studies of conflict more generally. Can we improve on these analyses—i.e., do our regressions fit the data better—if we replace the capability ratio with our new measure? To address this question, we replicate 18 recent analyses of conflict using DOE scores in place of the capability ratio or other functions of CINC scores. On the whole, we see that the models with DOE scores tend to have better in- and out-of-sample fit, though not always. In the remainder of this section, we describe the replication study and its findings, and we provide some guidance for selecting a measure in applied research.

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.[19] 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 (because of missing dyad-year identifiers), we excluded it from

[19] For details, see footnote 3.

		AIC			P.R.L.	
Replication	*N*	CINC	DOE	Vuong	CINC	DOE
Bennett (2006)	1,065,755†	29712	30969	−13.89	0.245	0.213
Weeks (2012)	766,272†	15816	15568	4.70	0.310	0.321
Jung (2014)	742,414†	10659	10588	2.44	0.350	0.354
Park and Colaresi (2014)	379,821†	10632	10587	2.68	0.315	0.318
Sobek, Abouharb and Ingram (2006)	183,451†	5344	5199	4.33	0.326	0.344
Gartzke (2007)	171,509†	4284	4167	4.04	0.442	0.457
Salehyan (2008*b*)	86,497	8864	8821	0.58	0.279	0.282
Fuhrmann and Sechser (2014)	85,306	2614	2580	1.36	0.203	0.208
Arena and Palmer (2009)	54,403†	1152	1061	2.77	0.071	0.137
Owsiak (2012)	15,806	5805	5750	2.36	0.117	0.125
Zawahri and Mitchell (2011)	12,186	814	809	0.66	0.062	0.066
Salehyan (2008*a*)	10,197	3003	2981	1.43	0.101	0.107
Fordham (2008)	7,788	537	604	−2.27	0.275	0.188
Dreyer (2010)	5,316	3676	3635	2.48	0.239	0.248
Huth, Croco and Appel (2012)	3,826	5938	5935	−0.64	0.053	0.052
Uzonyi, Souva and Golder (2012)	1,667	2008	1986	1.54	0.128	0.137
Weeks (2008)	1,276	1574	1568	0.03	0.101	0.105
Morrow (2007)	864	1488	1504	−2.81	0.260	0.251

Table 4. Summary of results from the replication analysis. In-sample goodness of fit is measured by the AIC and the Vuong (1989) test. 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 Vuong test statistic has a standard normal distribution under the null hypothesis of no difference between the models, so values with a magnitude of 1.96 or greater would lead us to reject the null hypothesis at the 0.05 significance level. Out-of-sample fit is measured by proportional reduction in log loss relative to the null model, as reported in the last two columns. We use repeated 10-fold cross-validation to estimate each model's out-of-sample log loss, with 10 repetitions for models indicated by a dagger (†) and 100 repetitions for all others. The null model's log loss is estimated via leave-one-out cross-validation.

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. In the end, we were left with the 18 studies listed in Table 4.

For each analysis in our sample, we begin by identifying 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. As a basic measure of each model's in-sample goodness of fit, we compute the Akaike (1974) Information Criterion,[21]

$$\text{AIC} = 2(\text{number of coefficients}) - 2(\text{log-likelihood}).$$

The AIC is commonly used in model selection, with lower values representing better fit. In addition, we compute the Vuong (1989) test of the null hypothesis that the original and replicated models fit equally well.[22] To estimate each model's out-of-sample fit, we perform repeated 10-fold cross-validation. Because each study has a discrete dependent variable, we again employ the log loss (equation (2)) to measure out-of-sample fit.

Table 4 summarizes the results of the replication analysis. In general, the models that include DOE scores do better than those with CINC scores by both in- and out-of-sample criteria. Starting with in-sample fit, we see that the DOE model has a lower AIC than the CINC model in 15 of 18 cases. Moreover, in more than half of those cases (8), under the Vuong test we would reject at the 0.05 significance level the null hypothesis that the models fit equally well. The difference in fit is also statistically significant in all three cases where the CINC model has a lower AIC. The results are similar for out-of-sample fit, with the DOE model having a greater proportional reduction in log loss in 14 cases. The improvement due to using DOE scores is typically modest—about a single percentage point increase in the proportional reduction in log loss.

[20] When no main model was apparent, our heuristic was 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] Because DOE scores are ternary, the replicated models typically have more parameters than their original counterparts, hence our use of the AIC (which penalizes over-parameterization) instead of the raw log-likelihood to measure in-sample fit.

[22] We employ the standard Bayesian Information Criterion (Schwarz 1978) correction to the Vuong test statistic.

With such a small sample of replicated studies, we can only conjecture about why DOE performs better in some cases and worse in others. We see that the cases where DOE is significantly better according to the Vuong test tend to have large sample sizes—but, then again, the study where it does worst has the largest N in our sample. In two of the replications where DOE performs worst, namely Bennett (2006) and Fordham (2008), we see that both specifications include the raw CINC scores alongside or in lieu of the capability ratio. These terms may be capturing monadic effects that the purely dyadic DOE scores miss. On the other hand, in the other three analyses that include raw CINC scores (Arena and Palmer 2009; Zawahri and Mitchell 2011; Weeks 2012), the replication with DOE scores performs better by both AIC and cross-validation loss.

Seeing as neither measure is uniformly better, how should empirical scholars choose which one to include in their analysis? Our first recommendation is to perform exactly the kind of analysis we have shown here—to compare the model fit using each measure according to criteria like the AIC, the Vuong test, or cross-validation. We stress that any model selection should be based on the overall fit of the model, which all three of the aforementioned statistics measure, and not how favorable each model is to the researcher's hypothesis. Second, theory also has a role to play. DOE scores directly measure each state's probability of winning a hypothetical international dispute; the capability ratio only represents raw capability shares, which we have seen are at best marginally related to expected dispute outcomes. When expectations are of primary interest, such as in tests of theories derived from the bargaining model of war (Fearon 1995), DOE scores should be preferred, all else equal. Conversely, if raw military capacity is of greater theoretical interest than expectations, researchers should lean toward including the capability ratio or other functions of CINC scores.

6 Conclusion

In this paper, we have argued that proxies should be constructed using predictive power as the criterion of interest, provided a method for doing so, and demonstrated the usefulness of the method in an application to measuring dispute outcomes. We hope that our efforts will be of use both for the DOE scores we provide and for the theoretical merits of our general argument.

In our application, 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 between two nations. It therefore has a more natural interpretation than the capability ratio. It also lacks the *ad hoc* assumptions imposed by both the CINC score and the ratio-based transformation used in most studies. 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. We hope, then, that it will find use as scholars advance and test new claims. And, in a substantive sense, we find it interesting that our more flexible approach suggests that material capabilities play an important role in determining dispute outcomes.

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—namely, the six Correlates of War National Material Capabilities variables. We did so consciously, as we wanted to demonstrate that our method could improve measures without introducing new covariates. Having made our point, we look forward to seeing what the future holds for coming versions of DOE when new data is brought to bear on the problem. At the risk of belaboring: we created DOE using open-source software and have made our replication code available, and so anybody with a computer—and some patience!—could create a new version with new covariates.

On the theoretical 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 $\boldsymbol{x}$ and some outcome of interest $\boldsymbol{y}$—the f we provide to map $\boldsymbol{x}$ to $\boldsymbol{y}$ will work. Just as with introducing new covariates in any given application, future scholars can improve their proxies by including new models for evaluation in the super learner—the general approach remains unchanged. Our application tasked us to create a proxy of a probabilistic expectation like those seen in formal models of choice under uncertainty, and similar applications provide a natural starting point for our method. Doing so, however, requires good theory for just what it is that we hope to predict with our abstractions—for example, what outcome could we use variables related to democracy, like those used in the Polity score (Marshall, Gurr and Jaggers 2014), to predict? We hope political scientists across subfields will turn their attention to examples like these as they construct new measures and improve existing ones.

We would like to conclude with a still broader point. Breiman (2001*b*) 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.[23] 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, many quantities to be proxied for are expectations, so they should be constructed with prediction in mind. 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 do not 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.

[23] In case it is not obvious from our previous citations, Breiman self-identifies as an algorithmic modeler. He claims that 98% of statisticians fall into the data modeling camp, or at least did as of 2001. We are comfortable positing that the percentage is similar, if not greater, for empirical political scientists in 2015.

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

All six variables are strongly 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:

$$h(x,\theta) = \sinh^{-1}(\theta x) = \log\left(\theta x + \sqrt{(\theta x)^2 + 1}\right). \tag{6}$$

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 5 gives the scale selected for each component. We use the transformed components in both the multiple imputation (see below) and the super learner training.

[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.

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 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

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

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

- 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 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 (5) via base R's `constrOptim()` function for optimization with linear constraints. 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

- 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*a*)

  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

  Estimator Heteroskedastic Probit

  CINC Terms CINC_A

DOE Terms p_A, p_B

Notes CINC and DOE terms are included in both the mean and dispersion equations.

- Bennett (2006)

  Model Replicated Table 1, Column 1

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

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

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

- Dreyer (2010)

  Model Replicated Table 2, Model 2

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

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

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

- Fordham (2008)

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

  Unit of Analysis Undirected Dyads

  Estimator Probit Regression

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

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

- Fuhrmann and Sechser (2014)

  Model Replicated Table 2, Model 1

  Unit of Analysis Directed Dyads

  Estimator Probit Regression

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

  DOE Terms p_A, p_B

- Gartzke (2007)

Model Replicated Table 1, Model 4

Unit of Analysis Undirected Dyads

Estimator Logistic Regression

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

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

- Huth, Croco and Appel (2012)

  Model Replicated Table 2

  Unit of Analysis Directed Dyads

  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

- Jung (2014)

  Model Replicated Table 1, Model 1

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

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

  DOE Terms p_A, p_B

- Morrow (2007)

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

  Unit of Analysis Directed Dyads

  Estimator Ordered Probit Regression

  CINC Terms $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$, interaction with joint ratification

  DOE Terms p_A, p_B, interactions of each with joint ratification

  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

Estimator Logistic Regression

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

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

- Park and Colaresi (2014)

  Model Replicated Table 1, Model 2

  Unit of Analysis Undirected Dyads

  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

- Salehyan (2008*a*)

  Model Replicated Table 1, Model 1

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

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

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

- Salehyan (2008*b*)

  Model Replicated Table 1, Model 2

  Unit of Analysis Directed Dyads

  Estimator Probit Regression

  CINC Terms $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$, interaction with refugee stock in A, interaction with refugee stock from A

  DOE Terms p_A, p_B, interaction of each with refugee stock in A, interaction of each with refugee stock from A

- Sobek, Abouharb and Ingram (2006)

  Model Replicated Table 1, first row (political prisoners model)

  Unit of Analysis Undirected Dyads

  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}

- Uzonyi, Souva and Golder (2012)

  Model Replicated Table 3, Model 3

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

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

  DOE Terms p_A, p_B

- Weeks (2008)

  Model Replicated Table 4, Model 3

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

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

  DOE Terms p_A, p_B

- Weeks (2012)

  Model Replicated Table 1, Model 2

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

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

  DOE Terms p_A, p_B

- Zawahri and Mitchell (2011)

  Model Replicated Table 2, Model 1

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

  CINC Terms CINC_A, CINC_B

  DOE Terms p_A, p_B

  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.

References

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

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.

Bafumi, Joseph, Andrew Gelman, David K. Park and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13(2):171–187.

Beardsley, Kyle. 2008. "Agreement without Peace? International Mediation and Time Inconsistency Problems." *American Journal of Political Science* 52(4):723–740.

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

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

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

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

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

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.

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.

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.

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.

Egesdal, Michael, Zhenyu Lai and Che-Lin Su. 2013. "Estimating Dynamic Discrete-Choice Games of Incomplete Information." Unpublished manuscript.

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

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro and Dinani Amorim. 2014. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" *The 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.

Gleditsch, Kristian S. and Michael D. Ward. 1999. "A Revised List of Independent States since the Congress of Vienna." *International Interactions* 25(4):393–413.

Hart, B.H. Liddell. 1972. *Why Don't We Learn from History?* London: George Allen & Unwin.

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

Heckelman, Jac C. and Keith L. Dougherty. 2013. "A Spatial Analysis of Delegate Voting at the Constitutional Convention." *Journal of Economic History* 73(2):407–444.

Hill, Daniel W. and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(03):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.

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/

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. 2001. "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference and Model Checking." *Political Analysis* 9(3):227–241.

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

Jacoby, William G. 2014. "Measurement in Political Science." Introduction to the virtual issue on measurement of *Political Analysis*.

James, Bill. 1983. *The Bill James Baseball Abstract, 1983*. New York: Ballantine.

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

Kadera, Kelly and Gerald Sorokin. 2004. "Measuring National Power." *International Interactions* 30(3):211–230.

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/

Kroenig, Matthew. 2009. "Exporting the Bomb: Why States Provide Sensitive Nuclear Assistance." *American Political Science Review* 103(1):113–133.

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

Kuhn, Thomas S. 1977. *The Essential Tradition: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Leeds, Brett Ashley. 2003. "Do Alliances Deter Aggression? The Influence of Military Alliances on the Initiation of Militarized Interstate Disputes." *American Journal of Political Science* 47(3):427–439.

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/

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

Maddala, G.S. 1977. *Econometrics*. New York: McGraw Hill.

Marshall, Monty G., Ted Robert Gurr and Keith Jaggers. 2014. "Polity IV Project: Dataset Users' Manual.".
URL: http://www.systemicpeace.org/inscr/p4manualv2013.pdf

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10(2):134–153.

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

McCallum, B.T. 1972. "Relative Asymptotic Bias from Errors of Omission and Measurement." *Econometrica* 40(4):757–758.

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.

Miller, Steven J. 2007. "A Derivation of the Pythagorean Won-Loss Formula in Baseball." *Chance* 20(1):40–48.

Mitchell, Sara McLaughlin and Paul R. Hensel. 2007. "International Institutions and Compliance with Agreements." *American Journal of Political Science* 51(4):721–737.

Molinaro, Annette M., Richard Simon and Ruth M. Pfeiffer. 2005. "Prediction Error Estimation: A Comparison of Resampling Methods." *Bioinformatics* 21(15):3301–3307.

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 Conflict1." *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.

Pitt, Mark A. and In Jae Myung. 2002. "When a Good Fit Can Be Bad." *Trends in Cognitive Sciences* 6(10):421–425.

Poole, Keith T. and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357–384.

Preacher, Kristopher J. 2006. "Quantifying Parsimony in Structural Equation Modeling." *Multivariate Behavioral Research* 41(3):227–259.

Quinlan, Ross. 2015. "Data Mining Tools See5 and C5.0.".
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.

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.

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

Shor, Boris and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105(3):530–551.

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.

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.

Spector, Paul E. 2006. Summated Rating Scale. In *The Sage Dictionary of Social Research Methods*, ed. Victor Jupp. London: Sage Publications pp. 295–297.

Stahlecker, Peter and Götz Trenkler. 1993. "Some Further Results on the Use of Proxy Variables in Prediction." *Review of Economics and Statistics* 75(4):707–711.

Su, Che-Lin and Kenneth L. Judd. 2012. "Constrained Optimization Approaches to Estimation of Structural Models." *Econometrica* 80(5):2213–2230.

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

Tibshirani, Robert and Guenther Walther. 2005. "Cluster Validation by Prediction Strength." *Journal of Computational and Graphical Statistics* 14(3):511–528.

Tibshirani, Ryan J. and Robert Tibshirani. 2009. "A Bias Correction for the Minimum Error Rate in Cross-Validation." *The 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.

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

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

Weeks, Jessica L. 2008. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization* 62(01):35–64.

Weeks, Jessica L. 2012. "Strongmen and Straw Men: Authoritarian Regimes and the Initiation of International Conflict." *American Political Science Review* 106(02):326–347.

Whang, Taehee, Elena V. McLean and Douglas W. Kuberski. 2013. "Coercion, Information, and the Success of Sanction Threats." *American Journal of Political Science* 57(1):65–81.

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.

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