Capability Ratios Predict Nothing*

Robert J. Carroll† Brenton Kenkel‡

July 21, 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 based on the train-validate-test paradigm from machine learning. 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 Zach Jones and Marc Ratkovic 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. 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.
1 Introduction

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, the discipline has 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 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 datasets 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.\(^2\) This is especially unfortunate given that many abstract quantities, particularly those based on formal models of decision under uncertainty, reflect expectations. However, the measurement of any quantity suffers when a data set’s unique peculiarities are assigned too much explanatory import.

\(^1\) 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.
Theoretical expectations present unique measurement challenges. 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 (like so many other empirical conflict scholars) “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.

Dispute expectations therefore put us in the middle of a measurement impasse. 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 for the hypothetical war in question, 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 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. Just like the leader, users of the bargaining model want the estimate of war outcomes that predicts best rather than the one that fights the last war.

We aim to do right by the leader. 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 sought to include a proxy for \( p \) in their analyses. Most scholars have followed Leeds’ lead and used ratios of CINC scores.\(^3\) CINC ratios are inappropriate as a proxy for expected dispute outcomes for a variety of reasons, including problems with the CINC function itself, \textit{ad hoc} parameterizations, and issues of functional form. What is more, when evaluated on predictive performance, we find that \textit{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 (1.2%) over a null model. Conversely, the measure we construct—DOE (Dispute Outcome Expectations) scores—improves out-of-sample predictive performance by 16.8%. It is notable that this is the case despite the fact that we use the same component variables from which the CINC score is constructed.

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 over the capability ratio and its cousins, we are not asking applied international relations scholars to give anything up. We replicated 18 recent empirical studies that utilized the capability ratio and

\(^3\) 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 \textit{American Political Science Review, American Journal of Political Science, Journal of Politics, International Organization, and International Studies Quarterly}. 
then replaced it with the DOE score. In 15 of the 18 replications, the DOE score improved both in- and out-of-sample goodness of fit. This means that DOE score is better than the capability ratio for four reasons: it comports better with scholars’ theories of international relations; it avoids the underfitting of *ad hoc* measures while avoiding the overfitting of traditional measurement models; it has a natural interpretation as the probability of each dispute outcome; and it usually performs better in the kinds of analyses most empirical international relations scholars care about.

The analysis proceeds in five 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 4 describes the data and methods we use to construct a new proxy for expected dispute outcomes, and section 5 provides the results of our replications. The final section addresses next steps and concludes.

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

---

*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).*

*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).*
“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....”

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 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 the train-validate-test approach common in machine learning. While traditional statistical approaches to measurement conform to maximization or minimization of relevant error or likelihood within the entire data set, we instead aim to optimize predictive performance. First, we split our data into a larger training set used for model fitting and selection and a smaller test set that we use to assess the predictive performance of our chosen model. We then fit a variety of models to the training set in order

---

6Our formal criterion for predictive performance, the log loss function, is explicitly described in the methods section.
to find the one with the best predictive performance. At this point, if we were to pick the model that best fits the training data, we would likely end up with one that is overfit. To prevent this, we assess model fit through cross-validation, which amounts to another layer of sample-splitting. Finally, once we have chosen a model from the training set, we apply it to the test set to yield an unbiased measure of its true predictive power when brought to previously unseen data.

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, interpretive nuance, and conservatism in inference—pale in comparison to these benefits.

3 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,

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

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

where $\alpha \geq 0$ adjusts $x$’s shape. Here the quest for the best-fitting $g$ amounts to estimating $\alpha$; 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

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

We now introduce the CINC function.\(^7\) State $i$’s CINC score in year $t$ is its average share across all variables:

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

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:

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

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

Many peculiarities 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

\[^7\text{Here CINC stands for “Composite Index of National Capability” as given in the Correlates of War National Material Capabilities data (Singer, Bremer and Stuckey 1972).}\]
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.

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 2008a) 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 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 Kubeski 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

---

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

$$f_{Bremer}(M_t) = \max \{CINC_a(M_t), CINC_a(M_t)\}$$

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.
decision to fight. So, while it remains true that the capability ratio is not meant to directly relate to $p$ in a bargaining model, the capability ratio was meant to tell us something about how uncertainly 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—might be a laudable first step, but it would still suffer from functional form dependencies. The careful scholar might sidestep these by fitting an ensemble of models with different functional forms imposed, but so long as the ensemble is fit to the entire capability data set, it will suffer from overfitting and lose predictive power. 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.

4 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 2001b). Therefore, to build our new measure, we augment traditional approaches with methods from machine learning.

4.1 Data

To evaluate the predictive performance of both the capability ratio and our alternative model, 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. Every dispute outcome is either A Wins, B Wins, or Stalemate, which we denote by $Y_i \in \{A, B, \emptyset\}$, respectively. In each dispute, we observe the year it took place, each disputant’s raw material capability components, and each disputant’s share of the system-wide total of each component at the time. These predictors are collected in the vector $X_i$.

---

9 See the Appendix for the data construction and coding specifics.
10 There are missing observations in the National Material Capabilities data. Consequently, about 17% of the disputes we observe contain at least one missing cell. We use multiple imputation to deal with missingness (Honaker and King 2010); see the Appendix for details.
We want to measure the out-of-sample predictive power of various models, which means we cannot use the full dataset for model fitting. Following common practice in machine learning, we randomly divide our sample into two parts: a training sample for building and selecting models, and a test sample for making out-of-sample predictions with a model built on the training sample. We use an 80-20 split, resulting in a training sample of $N_{\text{train}} = 1,391$ disputes and a test sample of $N_{\text{test}} = 349$ disputes.\footnote{To avoid complications due to missing data, we draw the test sample exclusively from the subset of complete observations.} We exclusively use the training sample for all model tuning, fitting, and selection. Even if we try to prevent overfitting in the training stage—which we do—we cannot rule out the possibility that the resulting model overfits the training data. This is where the test sample comes in. After selecting a model from the training data, we run the test sample through it to evaluate the model’s out-of-sample predictive power. As long as we do not use the test sample to make any further modeling decisions, this gives us an unbiased estimate of our chosen model’s out-of-sample performance.

### 4.2 Modeling Dispute Outcomes

Our goal is to build the model that squeezes the most out-of-sample predictive power out of the few predictors available to us.\footnote{Having more predictors would, of course, be even better—we just want to see how much better we can do than the capability ratio without any additional data collection. Our method for constructing a measure would generalize easily to a setting with additional predictors.} To begin with, we must be a bit more specific about what we mean by “predictive power.” We define a model as a function $f$ that maps from the predictor variables into the probability of each potential dispute outcome, $f(X_i) = (f_A(X_i), f_B(X_i), f_H(X_i))$. Our metric for a model’s predictive power is its log loss, which is common in multinomial classification settings and is closely related to the log-likelihood (Hastie, Tibshirani and Friedman 2009, 221). To calculate the log loss, we take a model’s predicted probability of each observed outcome, take the average of their logged values, and multiply the resulting average by $-1$. Smaller values of the log loss represent greater predictive accuracy, with a lower bound of 0 representing perfect prediction.\footnote{Predicted probabilities very close to 0 are trimmed to keep the log loss finite.}

Formally, the log loss of a model $f$ on the data $(X, Y)$ is

$$\ell(f, X, Y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \{A, B, H\}} 1\{Y_i = t\} \log f_t(X_i). \quad (2)$$

To select a model to use for out-of-sample prediction, we will fit a number...
of candidate models on the training data. Since we cannot use the test data for model selection, this requires estimating each model’s out-of-sample prediction error within the training sample. We cannot rely on ordinary within-sample measures of fit (e.g., percent correctly predicted), as these will lead us to overfit to the training data. Instead, we estimate each model’s out-of-sample log loss via $K$-fold cross-validation (Hastie, Tibshirani and Friedman 2009, 241–249). We randomly assign each observation in the training sample to a “fold” $k \in \{1, \ldots, K\}$ and then fit each candidate model $K$ times, each time leaving one fold out of the fitting.\footnote{When fitting models with tuning parameters, we choose tuning parameters separately within each fold via another cross-validation loop, again using log loss as the objective function. In this case, when choosing the tuning parameters for the model fit to the full training data, we use different folds than those we use to estimate its out-of-fold log loss.} Assume that we have $M$ candidate models indexed by $m = 1, \ldots, M$, and let $\hat{f}_m^{(-k)}$ be the result when we fit the $m$'th candidate model to the data excluding fold $k$. Our estimate of the out-of-sample log loss of the $m$'th candidate model is its average out-of-fold log loss,

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

where $(X^{(k)}, Y^{(k)})$ is the subset of training data assigned to fold $k$. Following usual practice, we use $K = 10$ cross-validation folds.

It is well known in the machine learning literature that averaging many models can lead to better predictive accuracy than using a single model (Breiman 1996). Accordingly, instead of simply choosing the model with the lowest CV loss, we combine the models in a weighted average following the super learner algorithm (van der Laan, Polley and Hubbard 2007). Specifically, we select the weights $\hat{w}_1, \ldots, \hat{w}_M$ that solve

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

where each $\hat{f}_m$ is the $m$'th candidate model fit to the full training data. Our final model is the super learner, $\hat{f} = \sum_m \hat{w}_m \hat{f}_m$. By definition, the CV loss of the super learner is no greater than that of the best candidate model, which is a special case.

To extract as much predictive power as possible from the material capability data, we examine a diverse array of candidate models to plug into the super learner.
learner. Each candidate model must, of course, work with a categorical outcome variable and generate predicted probabilities. Beyond that, we restrict our focus to classes of models that have been well studied in the machine learning or statistics literatures, settling on six: ordered logistic regression (McKelvey and Zavoina 1975), $k$-nearest neighbors (Cover and Hart 1967), random forests (Breiman 2001a), neural networks (Ripley 1996), Gaussian processes (Rasmussen and Williams 2006), and support vector machines (Cortes and Vapnik 1995). We run each model on four sets of variables: the disputants’ raw capability components and their capability component proportions, each with and without the year the dispute began. To mimic the applied literature, we include a standard ordered logistic regression on the capability ratio. Finally, we include a null model (ordered logistic regression on just an intercept), giving us $M = 26$ candidate models.

The biggest downside of our approach is that the results are not easily interpretable. Because the super learner entails averaging 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. Whether this is a problem depends on one’s 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.

4.3 Results

The analysis proceeds in two steps. First, we fit candidate models on the training set, cross-validate them to estimate their out-of-sample prediction error, and average them to form a super learner. Second, we compare the super learner’s predictions to actual outcomes in the test set to obtain an unbiased estimate of its out-of-sample predictive power. Our main findings are that the capability ratio has barely more predictive power than a null model and that the super learner does much better. In other words, by making optimal use of the component variables underlying the capability ratio, we can construct a superior proxy for expected dispute outcomes.

A glance at the training set results shows that the predictive power of the capability ratio is limited. We say “limited” not as a euphemism for “bad” (though it is bad), but rather literally. To see why, examine Table 1, which contains the results of the ordered logistic regression on the logged capability ratio in the training data. Although the coefficient on the capability ratio is positive and
<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capability Ratio (logged)</td>
<td>0.24</td>
<td>0.07</td>
<td>3.60</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Cutpoint: B Wins to Stalemate</td>
<td>−3.28</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutpoint: Stalemate to A Wins</td>
<td>1.76</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. 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.

statistically significant at the conventional level, its magnitude is small relative to the cutpoints. In order to predict that A Wins, we would need a logged capability ratio of more than 7—an impossibility, since the logged capability ratio is bounded above by 0. Similarly, to predict that B Wins, we would need a logged capability ratio of less than −13, which is below the minimal value observed in the training data (about −9). The predicted probabilities are also constricted: for example, the probability that A Wins is between 2 percent and 14 percent for all observations in the training data. So the capability ratio only weakly predicts dispute outcomes in sample. This does not bode well for its ability to make predictions out of sample.

Indeed, the results are even less encouraging when we turn to the cross-validation results, which give us a formal estimate of the capability ratio’s out-of-sample predictive power. The results for all candidate models, as well as their weights in the super learner ensemble, appear in Table 2. By the criterion of CV loss, the capability ratio does barely better than a null model at predicting dispute outcomes. Its proportional reduction in loss, a rough analog of $R^2$ for categorical data, is about 1 percent—a negligible improvement. Consequently, it is no surprise that the capability ratio model receives little weight in the optimal ensemble.\(^\text{15}\) Simply put, the capability ratio’s relationship with dispute outcomes is too weak for it to serve as a good proxy for the likelihood that a conflict will end in either side’s favor.

On the bright side, when we examine the other candidate models, we see that it is possible to build a better proxy with the same underlying components. All but two of the other candidate models have better cross-validation loss than the capability ratio model,\(^\text{16}\) and most have a proportional reduction in loss of 10 percent or greater. Looking at the performance figures reported in Table 2, a

\(^\text{15}\) Its average weight across imputations is about $7.5 \times 10^{-7}$.

\(^\text{16}\) The two models that perform worse are the Gaussian processes on the capability proportions. We suspect that the strong right skew of the capability proportions throws off the algorithm used for automatic tuning parameter selection in the software we use for Gaussian processes. See the Appendix for details.
<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Year</th>
<th>CV Loss</th>
<th>PR.L.</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Model</td>
<td>Intercept only</td>
<td>0.56</td>
<td></td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Ordered Logit</td>
<td>Capability Ratio</td>
<td>0.55</td>
<td>0.01</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Ordered Logit</td>
<td>Components</td>
<td>0.50</td>
<td>0.10</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Ordered Logit</td>
<td>Components ✓</td>
<td>0.50</td>
<td>0.09</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Ordered Logit</td>
<td>Proportions</td>
<td>0.54</td>
<td>0.04</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Ordered Logit</td>
<td>Proportions ✓</td>
<td>0.51</td>
<td>0.08</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>k-Nearest Neighbors</td>
<td>Components</td>
<td>0.49</td>
<td>0.13</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>k-Nearest Neighbors</td>
<td>Components ✓</td>
<td>0.46</td>
<td>0.17</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>k-Nearest Neighbors</td>
<td>Proportions</td>
<td>0.52</td>
<td>0.06</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>k-Nearest Neighbors</td>
<td>Proportions ✓</td>
<td>0.48</td>
<td>0.14</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>Components</td>
<td>0.49</td>
<td>0.12</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>Components ✓</td>
<td>0.48</td>
<td>0.14</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>Proportions</td>
<td>0.48</td>
<td>0.13</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>Proportions ✓</td>
<td>0.48</td>
<td>0.15</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Neural Network</td>
<td>Components</td>
<td>0.46</td>
<td>0.17</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Neural Network</td>
<td>Components ✓</td>
<td>0.47</td>
<td>0.16</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Neural Network</td>
<td>Proportions</td>
<td>0.50</td>
<td>0.10</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Neural Network</td>
<td>Proportions ✓</td>
<td>0.44</td>
<td>0.20</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Gaussian Process</td>
<td>Components</td>
<td>0.46</td>
<td>0.18</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Gaussian Process</td>
<td>Components ✓</td>
<td>0.45</td>
<td>0.19</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Gaussian Process</td>
<td>Proportions</td>
<td>0.66</td>
<td>−0.18</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Gaussian Process</td>
<td>Proportions ✓</td>
<td>0.56</td>
<td>−0.01</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Components</td>
<td>0.47</td>
<td>0.16</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Components ✓</td>
<td>0.48</td>
<td>0.15</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Proportions</td>
<td>0.50</td>
<td>0.10</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Proportions ✓</td>
<td>0.50</td>
<td>0.10</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Super Learner</td>
<td></td>
<td>0.43</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Summary of training set results, including the cross-validation estimate of out-of-sample log loss, proportional reduction in loss (relative to the null model), and optimal super learner weight of each candidate model. All quantities represent the average across imputed datasets.
Table 3. Results of applying the null model, the capability ratio model, and the super learner to the test data. PR.L. is the proportional reduction in log loss compared to the null model. Accuracy and kappa are measures of classification performance, where we take the predicted outcome to be the one with the highest predicted probability. Accuracy is the percentage correctly predicted, while kappa is the percentage improvement in classification over what would be expected by chance (Carletta 1996).

<table>
<thead>
<tr>
<th>Model</th>
<th>Log Loss</th>
<th>PR.L.</th>
<th>Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Model</td>
<td>0.458</td>
<td>0.880</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Capability Ratio</td>
<td>0.453</td>
<td>0.012</td>
<td>0.880</td>
<td>0.000</td>
</tr>
<tr>
<td>Super Learner</td>
<td>0.381</td>
<td>0.168</td>
<td>0.888</td>
<td>0.152</td>
</tr>
</tbody>
</table>

Few basic patterns emerge. Models that include the year of the dispute tend to predict outcomes better than those that do not, indicating that the importance of different material capability components varies over time. We also see, perhaps counterintuitively (and contrary to how the capability ratio is constructed), that the raw capability components tend to be better predictors than the annual totals. On the modeling front, neural networks tend to perform best, followed by random forests, support vector machines, and $k$-nearest neighbors. The ordered logistic regression models—significantly less flexible in terms of allowing for nonlinearities and interactions—perform worse.

As we expected, the super learner ensemble performs discernibly better than any of the candidate models from which it is constructed. The ensemble’s proportional reduction in loss, as estimated by cross-validation on the training set, is about 23 percent, or three percentage points better than the best candidate model. The model concentrates the bulk of its weight on two of the candidate models: the random forest on the raw capability components and the neural network on the annual capability proportions, both with year of dispute included. Interestingly, though one of these (the neural network) is the best individual model, the other is not the second-best. More generally, while models with lower CV loss 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.

Armed with these pleasant results, we proceed with the super learner to construct our proxy for expected dispute outcomes. Since cross-validation estimates of out-of-sample predictive power tend to be too optimistic (Tibshirani...
and Tibshirani 2009), we obtain an unbiased estimate by calculating the super learner’s log loss on our test sample. We reiterate here that the super learner is constructed solely from the training data and that we settled on its use before examining its test set performance. Had we done otherwise, the test set results might also be too optimistic in expectation. The results of the test set analysis appear in Table 3. The super learner’s out-of-sample proportional reduction in error is about 17 percent—less than our cross-validated estimate, as expected, but still respectable. Of course, we would like to have a model that does even better at forecasting dispute outcomes out of sample. We suspect this will take additional data on material capabilities or other factors that affect conflict processes. Given the diversity and flexibility of the models we apply to the National Material Capabilities data, it is hard to see how any further improvements would be more than marginal.

We also use the test data to confirm our pessimism about the capability ratio’s out-of-sample predictive power. We see from Table 3 that its proportional reduction in loss on the test sample is 1 percent, the same as what we estimated via cross-validation on the training sample. Beyond that, the test set results nicely illustrate the limited predictive range of the capability ratio model, as shown in the ternary plots in Figure 1. Under the capability ratio model, plotted in the left column, every dispute in the test set is given an 85–90 percent chance of being a stalemate. Seeing how narrow the capability ratio’s predictive range is, it is little surprise that it barely does better than a null model at prediction. Conversely, the super learner (depicted in the right column) 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. If our goal is to forecast the expected outcomes of potential international disputes, the super learner gives us a far better proxy than the capability ratio does.

4.4 The New Measure

With the super learner results in hand, we construct a new proxy for expected dispute outcomes. Like all of our candidate models, the trained super learner is a function that maps from the observed capabilities of a pair of states into the probability that a dispute between them will end in the initiator winning, the target winning, or a stalemate. 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
Figure 1. Predicted probabilities of dispute outcomes in the test set according to the capability ratio model and the super learner. The first row shows predictions for every dispute; the second excludes disputes that ended in a stalemate.
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.\textsuperscript{17}

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 relationship between the capability ratio and DOE scores is weaker than we expected. Across directed dyad–years, Country A's capability ratio is correlated with the DOE estimate of the probability A Wins at 0.10 and with B Wins at −0.32.\textsuperscript{18} As a close look at how the two measures might differ in particularly important cases, Figure 2 plots both measures for pairings of five major powers between 1860 and 2007. The only country for which the capability ratio and DOE scores consistently tell the same story—namely, one of decline on the in-

\textsuperscript{17} About 19 percent of directed dyad–years contain missing values of the capability components for at least one country. We average across multiple imputations of the capabilities data to calculate the DOE scores for these cases. See the Appendix for details.

\textsuperscript{18} The two correlations are not the inverse of each other since DOE scores are ternary, with Stalemate as the third category.
Figure 2. Comparison of the capability ratio to DOE estimates for selected major power dyads between 1860 and 2007. These plots use the undirected form of the DOE estimates.
ternational stage—is the United Kingdom. The two measures also track each other reasonably well in the United States–Russia dyad, with the US chance of victory taking a dip during the Cold War and increasing thereafter. Notice, though, that the US's capability ratio relative to Russia's increases sharply after the Cold War, whereas the DOE estimate of its chance of victory only goes up marginally. We see perhaps the most divergence in the China–Japan dyad, where increases in Japan's capability ratio seem to be associated with declines in the DOE prediction that it would prevail in a dispute. 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

We have introduced a new measure of expected dispute outcomes and shown that it is a noticeably better proxy than the capability ratio. The next question that arises is whether, or in what situations, the new measure is useful for the empirical study of international relations. The capability ratio belongs to the standard battery of control variables in statistical analyses of conflict. Can we improve on these analyses—i.e., do they 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.\footnote{\textsuperscript{19} For details, see footnote 3. In future iterations of the project, we are planning to add replications of analyses from the \textit{Journal of Conflict Resolution}.} 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 the analysis. We also excluded studies that employed duration models or selection models, due to conceptual and technical problems with assessing their out-of-
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.

<table>
<thead>
<tr>
<th>Replication</th>
<th>N</th>
<th>CINC</th>
<th>DOE</th>
<th>Vuong</th>
<th>CINC</th>
<th>DOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bennett (2006)</td>
<td>1,065,755†</td>
<td>29712</td>
<td>30836</td>
<td>−12.11</td>
<td>0.245</td>
<td>0.217</td>
</tr>
<tr>
<td>Weeks (2012)</td>
<td>766,272†</td>
<td>15816</td>
<td>15579</td>
<td>4.52</td>
<td>0.310</td>
<td>0.320</td>
</tr>
<tr>
<td>Jung (2014)</td>
<td>742,414†</td>
<td>10659</td>
<td>10569</td>
<td>2.82</td>
<td>0.350</td>
<td>0.355</td>
</tr>
<tr>
<td>Park and Colaresi (2014)</td>
<td>379,821†</td>
<td>10632</td>
<td>10589</td>
<td>2.46</td>
<td>0.315</td>
<td>0.318</td>
</tr>
<tr>
<td>Sobek, Abouharb and Ingram (2006)</td>
<td>183,451†</td>
<td>5344</td>
<td>5232</td>
<td>3.69</td>
<td>0.326</td>
<td>0.340</td>
</tr>
<tr>
<td>Gartzke (2007)</td>
<td>171,509†</td>
<td>4284</td>
<td>4196</td>
<td>3.60</td>
<td>0.442</td>
<td>0.454</td>
</tr>
<tr>
<td>Salehyan (2008b)</td>
<td>86,497</td>
<td>8864</td>
<td>8840</td>
<td>−0.18</td>
<td>0.279</td>
<td>0.280</td>
</tr>
<tr>
<td>Fuhrmann and Sechser (2014)</td>
<td>85,306</td>
<td>2614</td>
<td>2569</td>
<td>1.63</td>
<td>0.203</td>
<td>0.211</td>
</tr>
<tr>
<td>Arena and Palmer (2009)</td>
<td>54,403†</td>
<td>1152</td>
<td>1120</td>
<td>0.69</td>
<td>0.071</td>
<td>0.086</td>
</tr>
<tr>
<td>Owsiak (2012)</td>
<td>15,806</td>
<td>5805</td>
<td>5679</td>
<td>4.46</td>
<td>0.117</td>
<td>0.136</td>
</tr>
<tr>
<td>Zawahri and Mitchell (2011)</td>
<td>12,186</td>
<td>814</td>
<td>806</td>
<td>0.85</td>
<td>0.062</td>
<td>0.068</td>
</tr>
<tr>
<td>Salehyan (2008a)</td>
<td>10,197</td>
<td>3003</td>
<td>2993</td>
<td>0.43</td>
<td>0.101</td>
<td>0.104</td>
</tr>
<tr>
<td>Fordham (2008)</td>
<td>7,788</td>
<td>537</td>
<td>650</td>
<td>−3.79</td>
<td>0.275</td>
<td>0.131</td>
</tr>
<tr>
<td>Dreyer (2010)</td>
<td>5,316</td>
<td>3676</td>
<td>3632</td>
<td>2.68</td>
<td>0.239</td>
<td>0.249</td>
</tr>
<tr>
<td>Huth, Croco and Appel (2012)</td>
<td>3,826</td>
<td>5938</td>
<td>5941</td>
<td>−1.11</td>
<td>0.053</td>
<td>0.051</td>
</tr>
<tr>
<td>Uzonyi, Souva and Golder (2012)</td>
<td>1,667</td>
<td>2008</td>
<td>1991</td>
<td>1.15</td>
<td>0.128</td>
<td>0.135</td>
</tr>
<tr>
<td>Weeks (2008)</td>
<td>1,276</td>
<td>1574</td>
<td>1570</td>
<td>−0.33</td>
<td>0.101</td>
<td>0.104</td>
</tr>
<tr>
<td>Morrow (2007)</td>
<td>864</td>
<td>1488</td>
<td>1478</td>
<td>−0.36</td>
<td>0.260</td>
<td>0.264</td>
</tr>
</tbody>
</table>
sample performance. Lastly, we excluded studies in which our measure of expected dispute outcomes would be endogenous, namely those whose dependent variable was the same as or closely related to the one we used to construct the DOE scores. 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. 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, \[ \text{AIC} = 2 \times (\text{number of coefficients}) - 2 \times (\text{log-likelihood}). \]

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. 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 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 about half of those cases (7), under the Vuong test we would reject the null hypothesis that the models fit equally well at the 0.05 significance level. The difference in fit is also statistically significant in two of the three cases where the CINC model fits the sample better. The results are similar for out-of-sample fit, with the DOE model having a greater proportional reduction in log loss in the same 15 cases. The improvement due to using DOE scores is typically modest—about a single percentage point increase in the proportional reduction in log loss.

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

---

\[20\] 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.
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. When we look at the two 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 assump-
tions 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.

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 $x$ and some outcome of interest $y$—the $f$ we provide to map $x$ to $y$ will work. Just as with introducing new covariates in any given application, future scholars can improve their proxies by including new models 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 (2001b) 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.\footnote{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.} 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.
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). 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. Table 5 lists the proportions of zeroes and missing values among each variable.

<table>
<thead>
<tr>
<th>Component</th>
<th>Pr(Zero)</th>
<th>Pr(Missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iron and Steel Production</td>
<td>0.558</td>
<td>0.006</td>
</tr>
<tr>
<td>Military Expenditures</td>
<td>0.034</td>
<td>0.139</td>
</tr>
<tr>
<td>Military Personnel</td>
<td>0.066</td>
<td>0.027</td>
</tr>
<tr>
<td>Primary Energy Consumption</td>
<td>0.097</td>
<td>0.030</td>
</tr>
<tr>
<td>Total Population</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>Urban Population</td>
<td>0.210</td>
<td>0.007</td>
</tr>
</tbody>
</table>

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, we transform each according to the inverse hyperbolic sine to correct for skewness (Burbidge, Magee and Robb 1988). We use the transformed components in both the multiple imputation (see below) and the super learner training.

A.2 Militarized Interstate Dispute Data

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

26 Downloaded from http://correlatesofwar.org/data-sets/MIDs/mid-level/at_download/file.
disputes that end in an outcome other than one side winning, one side yielding, or a stalemate;\textsuperscript{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. Table 6 lists the distribution of outcomes in the full data and in the split subsamples.

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Wins</td>
<td>201</td>
<td>174</td>
<td>27</td>
</tr>
<tr>
<td>Stalemate</td>
<td>1460</td>
<td>1153</td>
<td>307</td>
</tr>
<tr>
<td>B Wins</td>
<td>79</td>
<td>64</td>
<td>15</td>
</tr>
</tbody>
</table>

\textbf{Table 6.} Distribution of the three dispute outcomes in the full dataset, the training subset, and the test subset.

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.

\section*{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 (Honaker and King 2010). We perform the imputations via the \texttt{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. 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 \texttt{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.

\textsuperscript{27} For details on other kinds of outcomes, see the MID Codebook.

28
• 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 tcsPlot() function in Amelia, are available on request.

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 (4) via the constrOptim() function for optimization with linear constraints in base R. Further details about each candidate model are summarized below.
• Ordered Logistic Regression

**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

• *k*-Nearest Neighbors

**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

• Random Forest

**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\} for models without year and from \{2, 3, 5, \ldots, 13\} for models with year

**Notes** 1,000 trees per fit

• Neural Network

**Package** nnet (Venables and Ripley 2002)

**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^{-d}\}_{d=0}^{4}

**Notes** Single hidden layer, no skip layer, softmax fitting

• Gaussian Process

**Package** kernlab (Karatzoglou et al. 2004)

**Tuning Parameters**
Kernel width (sigma): selected automatically via `sigest()` in the `kernlab` package

**Notes** Radial basis kernel, all predictors centered and scaled to have zero mean and unit variance

- Support Vector Machine

**Package** `kernlab` (Karatzoglou et al. 2004)

**Tuning Parameters**

- Kernel width (sigma): selected automatically via `sigest()` in the `kernlab` package
- Constraint violation cost (C): selected via cross-validation from \( \{2^d\}_{d=-2}^{12} \)

**Notes** Radial basis kernel, all predictors centered and scaled to have zero mean and unit variance

### 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**  $\text{CINC}_A$, $\text{CINC}_B$, $\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}}$

**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**  $\log \text{CINC}_U$, $\log \text{CINC}_2$

  **DOE Terms**  $\log p_U$, $\log p_2$

- 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 \( \frac{\text{CINC}_A}{(\text{CINC}_A + \text{CINC}_B)} \)
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 \( \frac{\text{CINC}_A}{(\text{CINC}_A + \text{CINC}_B)}, \) 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 \( \log(\frac{\text{CINC}_\text{min}}{\text{CINC}_\text{max}}) \)
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 (2008a)

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 (2008b)

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 $p_{\text{min}}$, $p_{\text{max}}$

• Uzonyi, Souva and Gold (2012)

Model Replicated Table 3, Model 3
Unit of Analysis Directed Dyads
Estimator Logistic Regression
CINC Terms  \( \frac{CINC_A}{(CINC_A + CINC_B)} \)

DOE Terms  \( p_A, p_B \)

- Weeks (2008)

  Model Replicated  Table 4, Model 3
  Unit of Analysis  Directed Dyads
  Estimator  Logistic Regression
  CINC Terms  \( \frac{CINC_A}{(CINC_A + CINC_B)} \)
  DOE Terms  \( p_A, p_B \)

- Weeks (2012)

  Model Replicated  Table 1, Model 2
  Unit of Analysis  Directed Dyads
  Estimator  Logistic Regression
  CINC Terms  \( CINC_A, CINC_B, \frac{CINC_A}{(CINC_A + CINC_B)} \)
  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


**URL:** [http://CRAN.R-project.org/doc/Rnews/](http://CRAN.R-project.org/doc/Rnews/)


