Theory and Evidence in International Conflict: A Response to de Marchi, Gelpi, and Grynaviski

NATHANIEL BECK *New York University*
GARY KING *Harvard University*
LANGCHE ZENG *George Washington University*

In this article, we show that de Marchi, Gelpi, and Grynaviski's substantive analyses are fully consistent with our prior theoretical conjecture about international conflict. We note that they also agree with our main methodological point that out-of-sample forecasting performance should be a primary standard used to evaluate international conflict studies. However, we demonstrate that all other methodological conclusions drawn by de Marchi, Gelpi, and Gryanaviski are false. For example, by using the same evaluative criterion for both models, it is easy to see that their claim that properly specified logit models outperform neural network models is incorrect. Finally, we show that flexible neural network models are able to identify important empirical relationships between democracy and conflict that the logit model excludes a priori; this should not be surprising since the logit model is merely a limiting special case of the neural network model.

We thank Scott de Marchi, Christopher Gelpi, and Jeffrey Grynaviski (2004; hereafter dGG) for their careful attention to our work (Beck, King, and Zeng 2000; hereafter BKZ) and for raising some important methodological issues that we agree deserve readers' attention. We are pleased that dGG's analyses are consistent with the theoretical conjecture about international conflict put forward in BKZ—"The causes of conflict, theorized to be important but often found to be small or ephemeral, are indeed tiny for the vast majority of dyads, but they are large, stable, and replicable whenever the ex ante probability of conflict is large" (21)—and that dGG agree with our main methodological point, that out-of-sample forecasting performance should always be one of the standards used to judge studies of international conflict and, indeed, most other areas of political science.

However, dGG frequently err when they draw methodological conclusions. Their central claim involves the superiority of logit over neural network models for international conflict data, as judged by forecasting performance and other properties such as ease of use and interpretation ("neural networks hold few unambiguous advantages . . . and carry significant costs" relative to logit [dGG, p. 378]). We show here that this claim, which would be regarded as stunning in any of the diverse fields in which both methods are more commonly used, is false. We also show that dGG's methodological errors and the restrictive model they favor cause them to miss and mischaracterize crucial patterns in the causes of international conflict.

We begin in the next section by summarizing the growing support for our conjecture about international conflict. The subsequent section discusses the theoretical reasons why neural networks dominate logistic regression, correcting a number of methodological errors. The next section demonstrates empirically, with the same data as used in BKZ and dGG, that neural networks substantially outperform dGG's logit model. We show that neural networks improve on the forecasts from logit as much as logit improves on a model with no theoretical variables. We also show how dGG's logit analysis assumed, rather than estimated, the answer to the central question about the literature's most important finding, the effect of democracy on war. Because this and other substantive assumptions underlying their logit model are wrong, their substantive conclusion about the democratic peace is also wrong. The neural network models we used in BKZ not only avoid these difficulties, but they, or one of the other methods available that do not make highly restrictive assumptions about the exact functional form, are just what is called for to study the observable implications of our conjecture.

SUPPORT FOR BKZ'S CONJECTURE ABOUT INTERNATIONAL CONFLICT

The explanation of what drives international conflict put forward in BKZ (22) was built on a simple conjecture, that "the effects of most explanatory variables are undetectably small for the vast majority of dyads, but they are large, stable, and replicable when the ex ante probability of conflict is large." That is, our point is that

Nathaniel Beck is Professor, Department of Politics, New York University, New York, NY 10023 (Nathaniel.Beck@nyu.edu).

Gary King is David Florence Professor of Government, Center for Basic Research in the Social Sciences, Cambridge MA 02138 (http://GKing.Harvard.Edu; King@Harvard.Edu).

Langche Zeng is Associate Professor, Department of Political Science, The George Washington University, Washington, DC 20052 (lzeng@gwu.edu).

Many thanks to Scott de Marchi, Chris Gelpi, and Jeff Grynaviski for providing data and replication information and for useful comments on an earlier version of our paper and to Kyle Beardsley, Alexis Diamond, and Chad Rector for superb research assistance and comments on an earlier draft. Thanks also to Bruce Bueno de Mesquita, John Oneal, Bruce Russett, Phil Schrodt, Curt Signorino, and Randy Siverson for comments on an earlier draft. A replication data file is available at http://gking.harvard.edu/data.shtml#toe-resp. For research support, we thank the National Science Foundation (SES-0214027, SES-0112072, SES-0318275, and IIS-9874747), the Weatherhead Initiative, and the National Institutes of Aging (P01 AG17625-01).

theories of international conflict that have a one-size-fits-all approach to regions and time periods should be replaced with theories that reflect the highly contingent and context-dependent nature of the phenomenon. We showed that this simple idea accounts for many specific observable implications consistent with the evidence. For example, if our conjecture is correct, then statistical analyses using methods that specify nearly identical effects for all observations, such as the logit models that had previously dominated the literature, should reveal apparently small to nonexistent effects or effects that vary across specifications. Such is indeed the case in most of the literature. Second, when the effects in the subset of high *ex ante* probability of war dyads are sufficiently large, these effects would be robust across specifications, even using logit. This is indeed the case for relatively atheoretical but powerful variables such as time since the last war and contiguity. Third, our conjecture implies that small changes in the sets of dyads included in an analysis would lead to disproportionate effects on the results, which would appear to account for much of the instability of results across articles in the literature. Fourth, when combined with the strong priors that exist among scholars in international conflict studies, our conjecture would lead us to expect to see results that vary considerably from researcher to researcher, which is precisely what we see. Fifth, data subsetting practices in the literature, such as limiting analyses to politically relevant dyads, would be expected to generate selection effects that strengthen results compared to the full data set, but the results would still not be as large as plausibility checks or qualitative researchers suggest. All these implications are consistent with evidence in the literature.

Our conjecture is also strongly supported by the neural network analyses in BKZ, which estimated and revealed the nonlinearities and massive interactions directly, and by the new analyses in Lagazio and Russett (2002). dGG's model also unambiguously reveals their estimated relationships to be much flatter or nonexistent among the low *ex ante* probability of war dyads, just as our approach predicts. Of course, logistic regression is a much more limited procedure in terms of the types of empirical results that are possible for it to produce. In fact, the effect of an explanatory variable in the usual logit model, measured in terms of a small change in an explanatory variable on the probability of war (i.e., a derivative), is restricted to be between only zero and one-quarter of the respective coefficient, at a maximum, and much narrower for low-probability rare events; these probabilities will also vary inflexibly, and very little as the data change, from dyad to dyad. Hence, some of the results consistent with our conjecture were effectively assumed by dGG as part of their logit model specification, rather than estimated; these parts imply dGG's implicit theoretical agreement with our conjecture, hard coded in their choice of specification, rather than direct empirical support. The nonlinear portions of dGG's specification means that the derivative of the probability of war with respect to democracy can vary a good deal more than their other variables, and so provide a somewhat better test of our conjecture, although still limited in terms of the types of results it can model.

WHY NEURAL NETWORKS DOMINATE LOGIT: THEORY

In this section, we provide the theoretical reasons why neural networks dominate logit models for the analysis of international conflict data[1] and correct a variety of methodological errors in dGG. We show that neural networks do not make "the interpretation of predictors extraordinarily difficult" as dGG claim, they do not produce "inefficient estimates," and they generate no extra "uncertainty about causal relationships"; and their claims that neural networks "preclude hypothesis testing" and that "it is not possible to determine the sensitivity of findings to changes in other variables" and "impossible to test hypotheses about the magnitude and direction of a predictor's influence on the dependent variable" are also false. Neural networks in fact are quite standard statistical models requiring no new "epistemology" or theory of inference. We now clarify these and several other points.

Epistemology and Interpretability

dGG claim that neural networks "embody a different epistemological perspective" from logistic regression. This claim is false. Neural network models may have a strange name, but they embody no mystery and, as commonly used, are no harder to interpret than logit models. As BKZ demonstrate, neural networks are ordinary statistical models, just like logit or linear regression. They work completely within the standard Bayesian or likelihood theories of inference, just like logit.

To see these points, let Y_i be 1 if dyad i is at war and 0 if it is at peace, and denote X_i a set of explanatory variables. Then, as we describe in BKZ, a linear regression (or "linear probability") model can be written

$$\Pr(Y_i = 1) = \text{linear}(X_i),$$

where $\text{linear}(X_i) = X_i\beta$, and a logit model can be written

$$\Pr(Y_i = 1) = \text{logit}(\text{linear}(X_i)),$$

where $\text{logit}(a) = 1/(1 + e^{-a})$. The extra logit function makes logistic regression parameters harder to interpret, and indeed the vast majority of published articles in political science do not interpret the parameters of logit models directly. Most political scientists instead choose to compute predicted values, first differences, and other quantities of interest (King, Tomz, and Wittenberg 2000). Neural network models are simple generalizations, which can be written

$$\Pr(Y_i = 1) = \text{logit}(\text{linear}(\text{logit}(\text{linear}(X_i)))). \quad \textbf{(1)}$$

[1] We could also show why neural networks also dominate discriminant analysis, but because the latter is not commonly used in political science, dGG focus on logit, and there exist many good reasons to prefer logit to discriminant analysis, we consider only dGG's logit here.

Just as with logit models, political scientists who use neural network models do not interpret their coefficients directly and instead report predicted values, first differences, or other quantities of interest (Andreou and Zombanakis 2001; Bearce 2000; Beck, King, and Zeng 2000; Borisyuk et al. 2001; King and Zeng 2002; Lagazio and Russet 2002; Zeng 1999, 2000). We see no reason to think that a predicted probability of war is any harder to interpret, no matter how it is calculated. Neural networks are certainly less familiar to political scientists, but in our view the enormously important public policy issues at stake mean that the improved forecasting performance that comes from using these techniques that we demonstrate in BKZ and the reanalysis below should outweigh any inconvenience some researchers may have in learning new methods.

dGG are incorrect that one cannot compute uncertainty measures such as standard errors, confidence intervals, or hypothesis tests about quantities of interest in neural network models. The marginal effect plots in BKZ report error bars from a neural network analysis, which portray confidence intervals, and one can use standard Bayesian or frequentist theory to compute these for any quantity in the same manner as one computes it for logit. Of course, in any statistical model, hypothesis tests should not focus on arbitrarily parameterized coefficients but, rather, should be about quantities of real substantive interest to researchers.[2]

Model Flexibility

Although it has a parametric form that is almost as straightforward as logit, neural network models do have advantages over logit. They, but not logit models, have "arbitrary approximation capabilities" (White 1992). This means that at least one member of the neural network family of models (or a neural network model with a sufficient number of hidden neurons) can approximate any functional form suggested by the data, even if not specified by one's theory *ex ante*. This is a tremendous advantage, as it enables one to estimate relationships not known from prior theory. What logit models do in contrast is to make assumptions, some based on theory and some based on convenience, all of which require one to ignore empirical evidence. In contrast, neural network modeling is a more powerful information extraction tool. In fact, neural network models include logit models as limiting special cases, and so a proper use of neural network models should always outperform logit, at least in expectation (Hastie, Tibshirani, and Friedman 2001). Indeed, if the logit specification is correct, a competent neural network modeler should find that something arbitrarily close to the logit is the preferred specification. Moreover, both logit models and neural network models can be used for testing any relevant hypothesis; the only difference is that under logit the validity of these tests is conditional on a variety of more stringent assumptions.

The logit specification includes a whole range of restrictive assumptions that no prior theory or data in international conflict supports and no method other than those with flexible functional forms like neural networks is capable of testing. For example, the *only* functional forms to have been derived from formal theories of international conflict are massively violated by the restrictions of logit models, especially that logit probabilities are usually monotonic functions of the explanatory variables (Signorino 1999). Moreover, Signorino and Yilmaz (2003) prove that if even the simplest form of strategic interaction exists among the dyads, then the restrictions inherent in logit models make its estimates "biased and inconsistent." In contrast, neural network models can approximate and thus test models derived from strategic theory to any degree of precision. Thus, instead of using a statistical model with arbitrary restrictions, incapable of finding, testing, or confirming patterns indicated by theory, our general approach is to follow when feasible the simple maxim from King and Zeng (2002): "When we know something, we assume it; when we don't know, we estimate it."

The flexibility of neural networks contrasts vividly with the rigidity of a standard logit specification, which uses the *same* effect parameter when predicting the probability of war between Burkina Faso (one of the poorest countries in the world, located in western Africa) and St. Lucia (a small Caribbean island tourist destination) as it does for the probability of war between the United States and North Korea. (That is, the estimated β coefficient relating the change in the explanatory variables to the conflict outcome is the same for both dyads.) This strikes us as incorrect, if not absurd, but it is precisely the kind of assumption made by the vast majority of the scholarly literature prior to BKZ. It is also the assumption made by dGG's logit regressions. Given this point and the observable implications of our conjecture to the contrary offered in BKZ, we would expect to see effects that are highly variable across the dyads, with massive interaction effects and nonlinearities that could only partially be picked up by techniques like logistic regression.

dGG worry that neural network models "increase the size of the parameter space almost 30-fold." This is misleading. Neural networks typically have more parameters than logit models, but Bayesian regularization reduces this nominal number of parameters to an "effective number of parameters" that is usually much smaller (Bishop 1995, 377, 410). Of course, the number of parameters per se is not the right criterion by which to judge a model, for the complexity of a model should match that of the data. A model too simple to extract information from the data would be simply wrong and useless. Overfitting—modeling the idiosyncratic features of data rather than the systematic features that will persist—cannot be ascertained by naive parameter counts but rather requires understanding the demands being put on the data by the complete estimation procedure. Even the user's manual for the software dGG

[2] For our models reported in the text, we compute probabilities, as the building blocks for our quantities of interest, by integrating over the posterior distribution of the parameters. This incorporates all information on uncertainty in our final estimates, rendering signficance tests irrelevant and, for rare events data like wars, reducing mean square error (King and Zeng 2001).

use warns against naive parameter counting like this (Demuth and Beale 2002, 5–54). By the appropriate use of Bayesian prior densities—and most importantly the use of test sets during the training stage rather than merely in-sample t-tests—neural network models enable researchers to estimate relationships rather than assume them, and at the same time they can help avoid overfitting and the resultant suboptimal forecasts. A logit model used in the standard way, without test sets or cross-validation, has little to prevent it from overfitting, even though it is a much less flexible form.

Neural networks also avoid "underfitting," a key problem with logit analysis, which makes assumptions about fundamental substantive relationships that we know little about. In fields where prior knowledge is extensive, specific, and highly informative, assumptions like these may be appropriate. Describing the international conflict literature in this way would be a stretch.

dGG's specification decisions exacerbate this inherent problem with logit analysis, because they appear to implement a notion of "theory" that does not depend heavily on prior evidence. To BKZ's specification, dGG "added variables and nonlinear transformations" that they claim "the literature has established as central to any model of dispute initiation." They cite 10 previous studies in support of this new specification. Whereas distance between the countries and whether one is a major power are good additions that we are happy to add to our model, what dGG do not say is that none of these prior articles include dGG's specification of nonlinear transformations or anything close to it. In fact, no other published or unpublished work we could find, whether or not it was cited in dGG, used this specification. Even the most recent publications (Reuveny and Li 2003; Russett, Oneal, and Berbaum 2003) and the most recent working paper by one of the authors (Gelpi and Grieco 2000) chose more traditional specifications, very much unlike the one in dGG.

Consider democracy, which is the variable that has received the most attention in the conflict literature among those considered by dGG. The raw measures for most specifications begin with the Polity scores for each country in a dyad, coded −10 (autocracy) to 10 (full democracy). dGG then add 11 to each (changing the range to 1–21) and include the product of these two scores and the square of this product. They exclude the main effects of each, the difference between the countries, the minimum or maximum of the two, and the sum of the two, which is equivalent to assuming that the effects of these variables more common in the literature are exactly zero (see Oneal and Russett 1997). These are exceptionally strong theoretical assumptions about the causes of international conflict hard coded into their logit model. If the assumptions are incorrect, then dGG's logit model will yield incorrect results, as the model is not flexible enough to adjust. dGG justify this specification by appeal to "theory" but *not one* prior publication in the literature (including those cited by dGG) has used, examined, or even discussed this specification.

To be more specific, under the dGG scoring, even ignoring the quadratic term, the difference between a dyad with two democracies and a mixed dyad (one democracy and one autocracy) is *21 times larger* than the distance between a mixed dyad and two autocracies. For the squared interaction the same difference is 441 times larger. Because nothing in the logit functional form could correct for such a difference, and indeed no theory or empirical analysis in the literature supports anything like it, this cannot be considered a plausible specification. Of course, the problem is not much better in the two standard scoring schemes in use in the various logit analyses of conflict in the literature. Both the "minimum democracy score" and the "binary democracy variable" approaches treat the totally autocratic and mixed dyads as if they were identical. While one might find some complicated nonlinear transform of the two democracy measures that works in particular applications under logit or other nearly linear models, the only reasonable general approach would use a flexible functional form, such as a neural network.

Although dGG's specification does not rise to a reasonable notion of prior theory, we have no objection to it as one *possible* (reduced form) theory of international conflict. Indeed, inspiration for some of the nonlinear terms suggested by dGG could have been coded directly from the neural network-generated marginal effect plots in BKZ. Hence, to stack the deck as much as possible in dGG's favor, we adopt their specification for the logit model used for empirical comparisons in the rest of this article. (In our neural network model, we follow dGG and include only the basic explanatory variables without their hard-coded interactions and nonlinear terms.)

The Role of Parsimony in Political Science

Although neural networks do not necessarily have a larger "effective number of parameters" than logit models, they do have a more complex mathematical form. This worries dGG, who declare, as a criterion for comparing approaches, that "models should be parsimonious (King, Keohane, and Verba 1994)." We think that dGG misread King, Keohane, and Verba (1994, 20), who wrote,

> The principle of choosing theories that imply a simple world is a rule that clearly applies in situations where there is a high degree of certainty that the world is indeed simple. Scholars in physics seem to find parsimony appropriate, but those in biology often think of it as absurd. . . . We believe it is only occasionally appropriate. . . . We should never insist on parsimony as a general principle of designing theories, but it is useful in those situations where we have some knowledge of the simplicity of the world we are studying. Our point is that we do not advise researchers to seek parsimony as an essential good, since there seems little reason to adopt it unless we already know a lot about a subject.

Thus, one cannot claim that the simplicity of the logit model is an advantage unless that simplicity is consistent with the data, and no theory or empirical evidence indicates that it is. (A generation ago methodologists argued about logit versus a simple linear probability

model; the argument for the latter, which is no longer taken seriously in the literature, is akin to dGG's, that is, that one can easily read off effects in the linear probability model and that interpretation of logit coefficients is difficult. That debate is now clearly archaic.)

Instead of using parsimony, King, Keohane, and Verba (1994) argue that we should choose theories that "maximize leverage," by searching for as many observable implications of a theory as possible and testing them, regardless of how complicated the theory itself is. As it turns out, the central conjecture of BKZ is highly parsimonious and maximizes leverage by producing many observable implications consistent with it. In general, neural network models tend to be less parsimonious than logit models, but at least for international conflict data they maximize leverage by producing many more observable implications consistent with the data. Moreover, the standard practice of evaluating models by out-of-sample forecasting tests ensures that parsimonious models will be chosen only when the empirical evidence indicates that they should be chosen.

Out-of-Sample Evaluations

dGG agree with BKZ that using out-of-sample forecasting evaluations to maximize leverage and thus evaluate models is critical. However, dGG miss two critical consequences of this decision. The more important consequence is that out-of-sample evaluations can be used to guard automatically against overfitting. The use of in-sample statistics, such as those used by dGG, encourage researchers to "peek" at the data in choosing a model, generate post hoc explanations for apparent patterns, and thus fit idiosyncrasies in the data that do not generalize. This overfitting becomes apparent when researchers use such a model to predict out-of-sample, a situation where they are vulnerable to being proven wrong. Thus, if neural networks were overfitting the in-sample data, this would simply cause them to perform worse in the out-of-sample forecasting tests that we and dGG use. If neural networks provide better out-of-sample forecasts, it cannot be because they overfit the in-sample data. Thus, while our Bayesian regularization scheme guards against overfitting, the rigorous evaluation of models via out-of-sample forecasting performance further ensures that models that are overfit will not be chosen.[3]

The commitment to out-of-sample validation also shows that evaluating neural networks is as easy as for logit models. And whereas some may prefer to live in a simple world where one only asks if t-ratios exceed two, the commitment to evaluation by out-of-sample forecasting shows that such convenience is bought at a huge cost. To give one simple example of how hypothesis testing can be naive, the squared democracy variable in the dGG logit model has a substantial t-ratio of about -5. Thus by the conventional in-sample criterion, we would think that this variable is an important predictor of peace. However, when we compare the out-of-sample forecasting performance of the dGG logit with and that without this variable, we find that forecasting performance is slightly improved by *excluding* the variable! Hence, the huge t-statistic merely reflects *overfitting*, a fact we would not know without out-of-sample tests. The key point is that significance tests are conditional on the veracity of the model, whereas out-of-sample tests are conditional only on the assumption that the out-of-sample data are generated by the same process as the in-sample data.

Although dGG tout as one of the advantages of logit its ability to test statistical significance (p. 372), their claim is incorrect and this "advantage" is both nonexistent and misleading. It is nonexistent because hypothesis tests on any quantity of interest can be carried out in neural networks as straightforwardly as in logit, because neural networks, as we note earlier, work completely within the standard Bayesian or likelihood theories of inference.[4] It is misleading in that the standard in-sample "statistical significance" paradigm using either model can easily lead to the choice of inferior models, as we have seen from the example of the squared democracy variable in dGG's analysis. Thus, we see little reason in this context to rely on in-sample significance tests and instead encourage model comparison by out-of-sample forecasting performance. And as BKZ and our reanalysis below illustrate, evaluating out-of-sample forecasting performance of a model, whether it is specified as a logit or a neural network, is equally easy. We now turn to a comparison of the out-of-sample forecasting properties of dGG's chosen logit model with our preferred neural network model.

EMPIRICAL EVIDENCE

In the four parts in this section, we discuss appropriate data analysis procedures using neural networks, how to evaluate forecasting performance, empirical comparisons of forecasting performance between dGG's logit model and our neural network model, and how dGG's logit model misrepresents substantive information in dGG's data.

Appropriate Data Analysis Procedures

The particular neural network models used in dGG appear inappropriate to the task. dGG explain that they chose these models by maximizing "the area under the ROC curve in the training set" (p. 376), which in normal

[3] Out-of-sample forecasting and cross-validation in general are routinely used techniques in modern statistical modeling. For an early reference on this topic, see Stone 1974.

[4] Both logit and neural network coefficients can be subject to hypothesis tests, but doing so for logit is easier. However, the ease of this test comes with a set of assumptions about the nonexistence of interactions and the veracity of a particular functional form that have no basis in the theory or evidence given in the international conflict literature. Moreover, coefficients from neither model are quantities of interest and so should not be the subject of hypothesis tests. Testing genuine quantities of interest, such as the probability of conflict or first differences, is as easy in both methods, assuming of course that estimation uncertainty has not already been integrated out, which would make hypothesis tests irrelevant.

circumstances is almost guaranteed to result in overfitting. dGG also chose not to use the software we used in BKZ or one of the other sufficiently sophisticated neural network programs available; their alternative choice was not well suited to avoiding overfitting, especially given their model selection criterion.[5]

dGG did recognize the importance of including Bayesian regularization (i.e., prior densities). However, the regularization scheme in their software was not designed for predicting a dichotomous dependent variable like war. As the manual for their computer program (Neural Network Toolbox for Matlab [see Demuth and Beale 2002, 5–60]) warns, "Bayesian regularization implemented in the toolbox does not perform as well on pattern recognition problems [i.e., classification problems, in particular binary dependent variables] as it does on function approximation [i.e., continuous dependent variable] problems." This is because, among other things, their program is running least squares with a dichotomous dependent variable, as if they were running a regression, rather than using the correct likelihood for the problem—a practice political scientists have known was flawed and for the most part abandoned since the 1970s. This program, by using a single regularization parameter, also makes the incorrect assumption that all their parameters have identical variances in their prior distributions. Bishop (1995, 340–42) proves that this simplistic assumption is inconsistent with a model that has appropriate scaling properties. Thus, their choice of software with a flawed error function and regularization procedure, uncorrected by true out-of-sample tests, apparently led to the neural network they chose overfitting their data and forecasting suboptimally.

Another problem is that dGG do not follow the standard practice of normalizing their data prior to running their neural network. Normalization is a key part of preprocessing (to which Bishop 1995, chap. 8, devotes an entire chapter) and has long been recognized as an integral part of neural network modeling. It involves transforming variables to mean zero and unit standard deviation for estimation; after estimation the variables are transformed back to the scale of interest. Normalization thus also scales the parameter values, which is important for any nonlinear optimization procedure, and is especially important here as, without it, the Bayesian priors would be scaled incorrectly. (In this data set, normalization may be critical, as the scales of the variables in their original units differ by a factor of more than 12,000!) As Hastie, Tibshirani, and Friedman (2001, 358) explain, normalization "can have a large effect on the quality of the final solution." Normalization is recommended by the user's guide to dGG's chosen software, which even provides commands for doing it automatically (Demuth and Beale 2002, 5–61).

In our analysis, below, we use the same software recommended in BKZ, which does not have the shortcomings of the Neural Network Toolbox for Matlab (and, unlike Matlab and the Toolbox, also has the advantage of being free and open source). We use the same variables as in dGG's neural network models, which include the basic linear terms of the input variables (appropriately normalized), but, following dGG, exclude the extra nonlinear transformations and interactions in dGG's logit model.

As in dGG's logit analysis, we evaluate the performance of the neural network using two evaluation sets, one consisting of the last four years of data (1986–89) and the other a random subset of the pre-1986 data. As a test of forecasting the future, the post-1985 data set seems better, but performance in this context depends not only on picking up the underlying structure in the data, but also on the assumption common to all methods in this context—that the process generating the in-sample data is the same as that generating the out-of-sample data. The problem with the post-1985 test, then, is that it is just one test and what it shows will differ from other out-of-sample tests. As a test of the comparative performance of the two methods when the out-of-sample data are known to be generated by the same process as the in-sample data, however, the random subset is near-optimal. Thus, although we present all our results for both test sets, it is the random subset results on which we should focus.

We chose the neural network architecture by using the in-sample data only (i.e., pre-1986 data minus the random subset) and avoid overfitting via the appropriate regularization procedures and by the use of test sets created temporarily from within the training data, all following the procedures described in BKZ. And following King and Zeng (2002), we use committee methods to reduce variance and improve out-of-sample performance. Finally, as in BKZ, we examine the out-of-sample evaluation sets once, only after we arrived at our final model. It is these final results that appear in this paper.

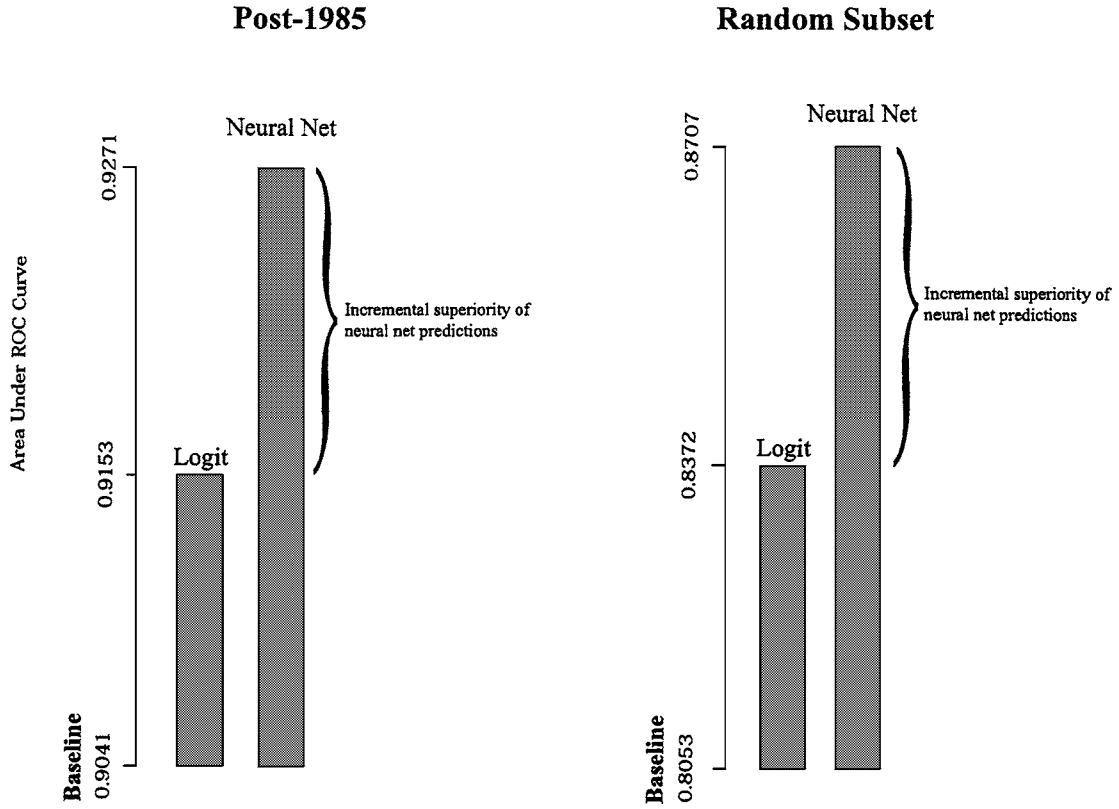
Evaluative Criteria

BKZ and dGG agree on the necessity of model evaluation via comparing out-of-sample forecasts but differ on the criteria used to summarize forecasting performance. BKZ's evaluative criterion was the number of zeros and (mostly) ones correctly predicted in out-of-sample test sets. Under this criterion, BKZ's neural network model outperformed BKZ's logit model. By BKZ's criterion, dGG's logit model *reduces* forecasting performance relative to BKZ's neural network and logit models, even though dGG's logit uses additional explanatory variables.

However, for new analyses, dGG argue that ROC curves are preferable to the evaluative criterion used in BKZ, and we are happy to go along. (ROC curves were first introduced in political science research by two of us after BKZ was published [King and Zeng 2002]; if we had been aware of them when we wrote BKZ, we obviously would have used them.) dGG's method of summarizing the information in ROC curves, by reporting the area under them, is not the only choice available, but it may be reasonable in this application

[5] That is, researchers who use test sets properly have much greater freedom in choosing software. A list of available neural network software can be found at http://gking.harvard.edu/nn.

FIGURE 1. Area under ROC Curves for Two Out-of-Sample Evaluation Sets



Note: The logit model was that run by dGG and represents, according to them, the "major findings of the quantitative security studies literature." The (atheoretical) baseline model is a logit analysis with only a constant term and a correction for duration dependence. Our neural network model indicates forecasting performance twice as accurate as dGG's logit model relative to the baseline.

and so we adopt it here too as the evaluative criterion. Obviously, a model chosen by optimizing according to one criterion will not necessarily do well on another, and evaluating it with respect to a criterion it was not designed to advance would make little sense. (By analogy, a least-squares estimator minimizes the sum of squared residuals and would not generally do well if evaluated by a minimum absolute deviations criterion.) Thus, whereas we chose a neural network model in BKZ by optimizing the number of zeros and (especially) ones correctly predicted, we now choose a model that optimizes according to ROC areas. For this paper, we do not estimate neural network models to satisfy other criteria.[6]

Forecasting Performance Comparison

We now compare three models using dGG's preferred criterion. First is dGG's logit model, which, with their gracious help, we were able to replicate exactly. Second is our neural network model. And, finally, is an "atheoretical baseline" model that has only a constant term and a correction for duration dependence (using dGG's splines based on the "years since the last war" variable) and merely says "peace persists." A miminal baseline model like this is of course a standard procedure used to facilitate comparison in a variety of fields.[7]

For each of the two evaluation sets, we calculated ROC curves (they appear in the Appendix) and then use dGG's preferred summary of them, the area under the ROC curve, for each of the three models. Figure 1 reports these results and conclusively demonstrates that the neural network model substantially outperforms dGG's logit model. The out-of-sample forecasting performance of the model that dGG characterize as

[6] Thus, the main differences in our neural network procedure from BKZ is using dGG's "ROC area" optimization criterion and explanatory variables. The only other difference is committee methods, which are increasingly used in neural network analyses and other areas due to their variance-reduction properties, which make the analysis more robust. Our conclusions would not change without this refinement because all members of the committee also outperform logit, but we recommend committee methods because it is now widely understood that they are normally superior. More specific details about our analysis are available in the replication data set accompanying this article. Our results for calibration tests, which are also used in dGG and BKZ, lead to similar conclusions as for ROC analyses, so for simplicity we only present the ROC analyses here.

[7] Our use of a baseline model is essentially the same as an economic forecaster who chooses a persistence model, that is, with a baseline forecast equal to the value of the variable during the last observed period. In contrast, a baseline equal to the global mean alone, or in our case assuming equal probabilities of war for all dyads, would ignore enormous time series dependence in the data and would therefore generate meaningless comparisons.

taking "into account the major findings of the quantitative security studies literature" (p. 375) is the leftmost bar in the figure. The height of this bar constitutes, in dGG's view, a summary of the sum total of the literature's prior knowledge about international conflict; the fact that dGG went to the extent of writing a paper to defend an entire "generation of scholarship" would seem to imply that they judge this performance to be substantial and substantively important. Some others will probably disagree (and we are happy to remain agnostic on this issue). But however one judges this performance, neural networks give us more than twice the forecasting accuracy and knowledge about the world, relative to the baseline, as all this prior knowledge about international conflict combined, relative to the same baseline. And this is without using any of the nonlinear terms used in dGG's logit model or adding a single extra datum or variable (and has no estimation uncertainty as the parameters are integrated out in the computation). The doubling of forecasting performance relative to baseline is approximately the same for the post-1985 test (on the left) as for the random subset of dyads (on the right).

Figure 1 was drawn using dGG's choice of a summary statistic, the ROC area, but to understand the numbers we need to translate them into units that are substantively meaningful. As is, the numbers are not comparable to numbers computed from other data sets and are not immediately related to any substantive feature of the results: We know that bigger is better but do not know whether the differences in ROC areas in the figure should be characterized as "large" or "small," and perusal of the full ROCs in the Appendix does not help either. Thus, restating the above more precisely, one meaningful unit is the distance from the atheoretical baseline to dGG's logit model, which according to dGG represents everything we know in the field, as summarized by dGG's logit. Using this as the unit of measurement, our neural network model takes us to more than two units. Thus, the marginal contribution of our methods—the distance between the performance of dGG's logit and our neural network—is slightly more than one unit. To assess whether a marginal improvement of one unit is "large" or "small" is thus equivalent to assessing whether the distance from the baseline to dGG's logit is large or small, as it is also about one unit. Because the one unit from the baseline to dGG's logit is dGG's summary of all the empirical knowledge existing in the field of conflict studies, dGG clearly imply that one unit is substantial. If this is correct, then our neural network's improvement over logit is just as substantial. Others may view the contribution of quantitative conflict studies as smaller than dGG do, and so should view the improvement of neural networks over logit to be concommitantly smaller as well, but even in this case the improvement is as large as in all prior research on the subject. We also show in the next section that this one incremental unit beyond dGG's logit turns out to be enormously important from a substantive perspective.

Another way to think about this is to consider the commonly reported proportionate reduction in error (PRE) statistic; the error in this case is one minus the ROC area (for comparison, the PRE in linear regression is R^2). For the post-1985 data, the PRE is 11.7% for logit and 24.0% for the neural network; for the random subset, the proportionate reduction in error is 16.4% for logit and 33.6% for the neural network.

Finally, as a robustness check on our results, we ran 4,400 additional neural network models based on different numbers of committee members, methods of committee decision making, procedures for calculuating the probability, network architectures, random seeds, and prior specifications—none of which were tuned according to the test procedures described in the text. Although skipping proper model selection rules is not advisable for purposes of inference or model comparison, these runs give a sense of how hard it would be to find a neural network model that forecasts worse than dGG's logit. It would be hard: Of the 4,400 runs, the logit was outperformed on the random test set (and in-sample data) by all 4,400 network models and on the post-1985 test set by 4,151 of the neural networks.[8]

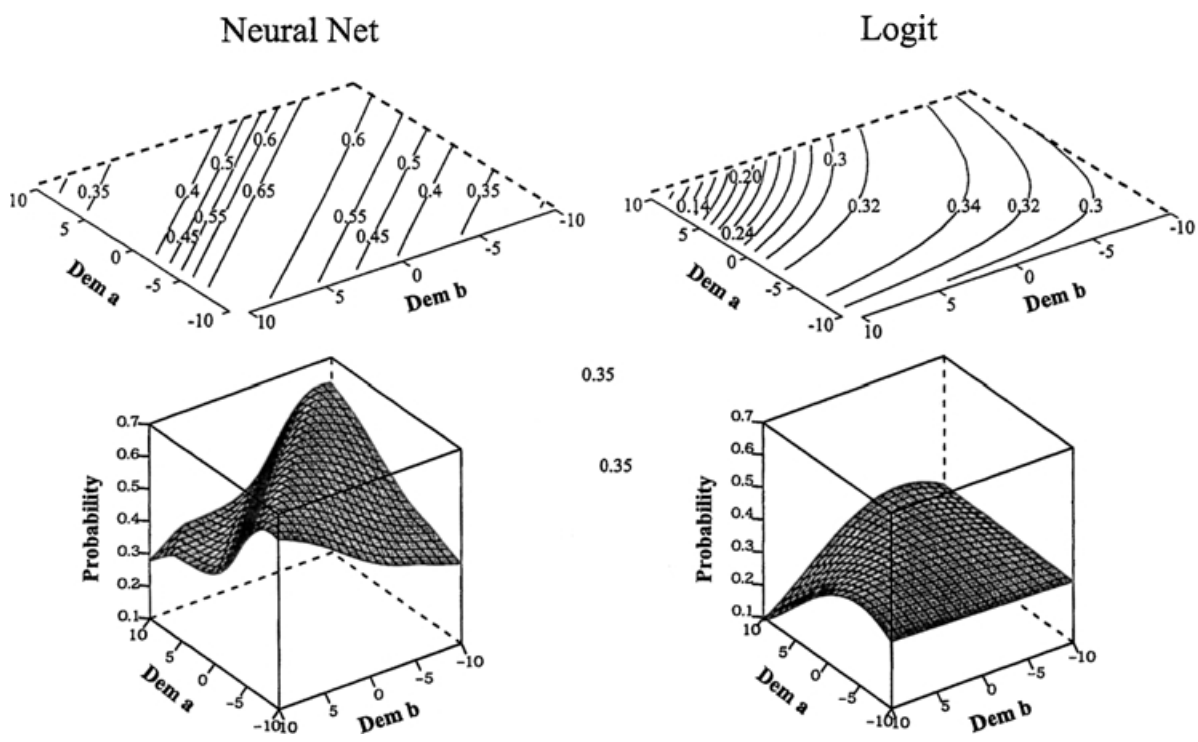
Substantive Implications

We now demonstrate that the substantive differences between dGG's logit specification and our neural network are of the utmost importance to the field. Indeed, dGG's logit specification included assumptions that virtually guaranteed that they would draw the wrong conclusions about the most important hypothesis in the literature, the effects of democracy on war, no matter what the data said. Many scholars have found that countries are less likely to fight each other if both are democratic; however, we do not know whether this is because they are democratic or because they are alike. In fact, there exists evidence in the literature that a pair of countries is less likely to fight if they are both autocracies under some circumstances (Peceny, Beer, and Sanchex-Terry 2002). A plausible alternative theory we now provide some evidence for is that "likes don't fight," in that dyads with different levels of democracy are those likely to go to war (Werner 2000). Because these issues remain highly controversial in the literature and remain the subject of a considerable research program, we think that dGG's specification—which makes some of these results impossible to obtain—unwisely substitutes unsupported theory for empirical analysis.

Figure 2 plots the probability of war as a function of the $[-10, 10]$ Polity score for the two countries in a dyad (marked "Dem a" and "Dem b"), holding constant other explanatory variables at values indicating high *ex ante* probabilities of conflict (computed from the median of the United States and China in their conflict years). We do this with a three-dimensional surface

[8] We also find, incidentally, that committee methods at least partially protect one from bias, in addition to reducing variance. The larger the committee, the less dependent results are to particular modeling choices and model selection procedures and the better are the forecasts on average. This result is unlikely to have arisen by chance, as it is consistent with findings from other fields in unrelated applications (e.g., Stock and Watson 2003).

FIGURE 2. Surface and Contour Plots for the Probability of War



Note: Surface and contour plots for the probability of war given differing levels of democracy (coded for each country in a dyad as −10 for autocratic to 10 for democratic) and a high *ex ante* probability of war. The logit model on the right is based on a replication of dGG's analysis. The neural network model on the left reveals features mischaracterized by the logit model.

plot (which makes it easy to visualize the relationship) and a contour plot (which helps provide numerical precision). The height of the surface is the probability of war. The contour plot should be thought of as the view of the surface plot from above, and the numbers reflect the probability for equal probability contours. For visual clarity we have not included confidence intervals, which are of no value for model selection with binary dependent variables anyway. dGG's logit model appears on the right: The dip in the surface at the left, where Dem a and Dem b are both near 10 (indicating full democracy), apparently supports the democratic peace hypothesis.

So far so good, until one realizes that other hypotheses were impossible to uncover with the logit model dGG specified, no matter how loudly the patterns in the data screamed. That is, the only results that could be found by dGG's specification were ones in which the probability of war was specified to be a logistic function of the product and the squared product of the degree of democracy of the two countries in each dyad (and other variables), a specification that mathematically restricts without theoretical reason the probability of conflict at $(-10, -10)$ to be similar to that at $(-10, 10)$ and $(10, -10)$. One way to demonstrate that the data were indeed screaming to get around this specification is by running our neural network model, as it encompasses dGG's logit model as a limiting special case. Hence, if the pattern on the right in Figure 2 were correct, the plots on the left in the figure, calculated from the neural network model, would display the same pattern. The differences, however, are massive. In the plot on the left, the low probability-of-war dyads are not uniquely clustered in the bottom-left corner where both countries are democratic. Instead, two clear regions of low probability of war appear, one reflecting the democratic peace and the other reflecting an autocratic peace. The ridge in the middle, indicating areas of high war probabilities, fit the "likes don't fight" hypothesis, as well as reflecting the well-known result that partial democracies (even if alike) tend to be more violent. In addition, the bottom-right corner, which was missed by logit, is not a minor feature of the data: The gradient from the top of the ridge to the right corner represents a remarkable 30 percentage point drop in probability—from 0.65 at the top to 0.35 at the right. This has obvious and critically important normative implications for the central hypothesis in the field.

Thus, for substantively oriented scholars of international conflict, neural network models are far more capable than logit models of representing the rich array of theoretical ideas in the literature and enabling scholars to draw valid empirical conclusions from existing data.

CONCLUDING REMARKS

dGG make three main points. We all clearly agree with their first point, that almost all researchers in this

field (and most other fields) should use out-of-sample forecasts, and tools like ROC and calibration curves, to evaluate statistical models. This will help end the practice of specifying models that are nearly invulnerable to being proven wrong and then drawing strong substantive conclusions without empirical grounding.

However, dGG's two other points are incorrect. A properly specified neural network model clearly outperforms dGG's logit model; in fact, a correctly constituted neural network model improves on the dGG out-of-sample forecast performance to the same degree that dGG improves on an atheoretical model that simply says that peace persists. dGG are also incorrect that neural network modeling involves some kind of novel epistemology or is of no interest to the theoretically minded student of international conflict. As we have seen, neural networks are interpreted using the same tools, in the same way as for logit, and both have the same statistical underpinnings. Neural network models, but not logit, are capable of fitting models where nations are acting strategically. Indeed, the logit model is simply a limiting special case of a neural network, and if the world is such that the simple logit form adequately represents the data, a correct neural network analysis will reveal this fact.

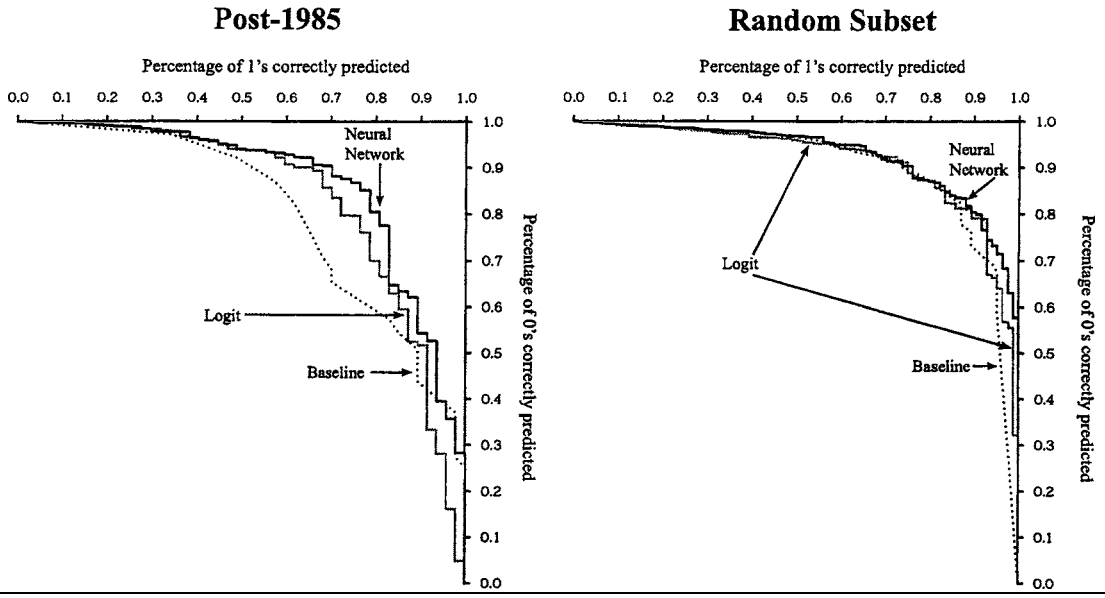
Not only does the neural network outperform logit for the data under consideration, but also we saw in Figure 2 that the neural network results are of tremendous interest for settling important controversies in the study of conflict. dGG's logit model is simply incapable of returning results consistent with an entire range of hypotheses in the literature, no matter what the data indicate. Clearly such a model cannot be used to decide fairly among competing hypotheses. Surely, data sets will exist where logit performs as well as neural networks but where neural networks (or other modern methods that enable one to estimate rather than assume the functional form) forecast better; we should neither fear to use them nor worry that our standard analytical methods must be discarded before we use them. Indeed, a whole range of methods has now been developed that does not require the functional form assumptions inherent in logit and the other techniques commonly used in political science. For prediction, these include neural networks, models of intermediate flexiblity like generalized additive models (Beck and Jackman 1998), those that have unique optima such as support vector machines (Vapnik 1995, 1998), and other types of models and methods such as boosting, regression and classification trees, kernel methods, and mixture models (Hastie, Tibshirani, and Friedman 2001). For estimating causal effects, in areas where pre- and posttreatment controls are clearly distinguishable, the techniques tend to be matching and related approaches (King and Zeng 2003). Which of these techniques is appropriate will depend on the application, but in almost all situations these techniques will dominate those with restrictive functional forms like logit, except in the presence of theory far more informative than in most areas of political science.

If our basic conjecture—that the world of international conflict combines a small number of dyads where a variable has a large effect with a large number of dyads where it has essentially no effect—is correct, then we must turn to appropriately flexible methods like neural networks that allow for these massive interactions that cannot be specified a priori. For the data and subject matter at hand, the massive interactions are there and logit is inadequate to deal with them.

APPENDIX: ROC CURVES

Figure 3 presents the three model ROC curves for each evaluation set. Clearly the ROC curve for the neural network

FIGURE 3. ROC Curves for the Post-1985 (Left) and Random (Right) Out-of-Sample Test Sets



Note: Each graph contains ROC curves for dGG's logit, our neural network, and the atheoretical baseline models.

model lies closer to the top right than the corresponding logit curve almost everywhere, and the gap between the curves is about as large as that between the baseline ROC and the logit curve.

Figure 1 provides a cleaner summary of the relevant characteristics of these graphs using dGG's preferred summary statistic, and the text explains how to understand the importance of the ROC differences seen here.

REFERENCES

Andreou, A. S., and G. A. Zombanakis. 2001. "A Neural Network Measurement of Relative Military Security—The Case of Greece and Cyprus." *Defence and Peace Economics* 12 (4): 303–24.

Bearce, David. 2000. "Economic Sanctions and Neural Networks: Forecasting Effectiveness and Reconsidering Cooperation." In *Political Complexity: Non Linear Models of Politics*, ed. Diana Richards. Ann Arbor: University of Michigan Press, 269–95.

Beck, Nathaniel, and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Model." *American Journal of Political Science* 42 (April): 596–627.

Beck, Nathaniel, Gray King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict." *American Political Science Review* 94 (March): 21–36.

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Borisyuk, Roman, Galina Borisyuk, Colin Rallings, and Michael Thrasher. 2001. "Forecasting the 2001 General Election Result: A Neural Network Approach." http://www.psa.ac.uk/spgrp/epop/forecasting_genelect2001.htm.

de Marchi, Scott, Christopher Gelpi, and Jeffrey D. Grynaviski. 2004. "Untangling Neural Nets." *American Political Science Review* 98 (2): 371–378.

Demuth, Howard, and Mark Beale. 2002. "Neural Network Toolboox for MATLAB, User's Guide." http://www.mathworks.com/.

Gelpi, Christopher, and Joseph M. Grieco. 2000. "Democracy, Interdependence, and the Liberal Peace." Duke University. http://www.duke.edu/~gelpi/papers.htm.

Hastie, Trevor, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning*. New York: Springer Verlag.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (Spring): 137–63.

King, Gary, and Langche Zeng. 2002. "Improving Forecasts of State Failure." *World Politics* 53 (July): 623–58.

King, Gary, and Langche Zeng. 2003. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." Preprint available at http://gking.harvard.edu.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (April): 341–55. Reprint at http://gking.harvard.edu/files/abs/making-abs.shtml.

Lagazio, Monica, and Bruce Russett. 2002. "A Neural Network Analysis of Militarized International Disputes, 1885–1992: Temporal Stability and Causal Complexity." In *The Scourge of War: New Extensions on an Old Problem*, ed. Paul Diehl. Ann Arbor: University of Michigan Press.

Oneal, John R., and Bruce Russett. 1997. "The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985." *International Studies Quarterly* 41 (June): 267–93.

Peceny, Mark, Caroline C. Beer, and Shannon Sanchez-Terry. 2002. "Dictatorial Peace?" *American Political Science Review* 96 (1): 15–26.

Reuveny, Rafael, and Quan Li. 2003. "The Joint Democracy-Dyadic Conflict Nexus: A Simultaneous Equations Model." *International Studies Quarterly* 47 (September): 325–46.

Russett, Bruce, John Oneal, and Michael L. Berbaum. 2003. "Causes of Peace: Democracy, Interdependence, and International Organizations, 1885–1992." *International Studies Quarterly* 47 (September): 371–93.

Signorino, Curtis. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93 (2): 279–98.

Signorino, Curtis S., and Kuzey Yilmaz. 2003. "Strategic Misspecification in Regression Models." *American Journal of Political Science* 47 (July): 551–66. http://www.rochester.edu/College/PSC/signorino/papers/Signo00.pdf.

Stock, James H., and Mark W. Watson. 2003. "Forecasting Output and Inflation: The Role of Asset Prices." *Journal of Economic Literature* 41 (September): 788–829.

Stone, M. 1974. "Cross-Validatory Choice and Assessment of Statistical Prediction." *Journal of the Royal Statistical Society, B* 36 (2): 111–33.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory.* New York: Springer.

Vapnik, Valdimir N. 1998. *Statistical Learning Theory.* New York: Wiley.

Werner, Suzanne. 2000. "The Effects of Political Similarity on the Onset of Militarized Disputes, 1816–1985." *Political Research Quarterly* 53 (June): 343–74.

White, Halbert H. 1992. *Artificial Neural Networks, Approximation and Learning Theory.* Cambridge, MA: Blackwell.

Zeng, Langche. 1999. "Classification and Prediction with Neural Network Models." *Sociological Methods and Research* 27 (May): 499–524.

Zeng, Langche. 2000. "Neural Network Models for Political Analysis." In *Political Complexity: Nonlinear Models of Politics*, ed. Diana Richards. Ann Arbor: University of Michigan Press, 239–268.