Nonparametric Model Discrimination in International Relations

KEVIN A. CLARKE
Department of Political Science
University of Rochester

This study introduces a simple nonparametric test for the relative discrimination of models in international relations research. The common parametric approach, the Vuong test, does not perform well under the small-*n*, high canonical correlation conditions that are sometimes encountered in world politics research. The nonparametric approach outperforms the Vuong test in Monte Carlo experiments and is trivial to implement even for the most complicated models. The method is applied to two empirical examples: the debate over long cycles and the effect of domestic politics on foreign policy decision making.

Keywords: *nonnested; nonparametric; model testing; Vuong*

International relations is a particularly paradigmatic area of study. Unlike the American subfield, which is characterized by strong theory and a high degree of conceptual agreement, international relations is rife with long-running battles between realists and nonrealists, general war theorists, long-cycle theorists, and theorists of the democratic peace.

This diversity of theoretical orientation is only partly the result of weak theory. A lack of adequate methods to compare theories empirically is a significant contributing factor. Comparing different theoretical positions, however, often requires comparing nonnested models.[1] Unfortunately, testing nonnested models outside the context of linear regression is often difficult, and existing methods are generally beyond the skillset of substantively oriented scholars.

To address this issue, I propose a simple nonparametric test for relative model discrimination. This new test, unlike existing parametric tests, is trivial to implement with any mainstream statistical software package. In addition, the new test, despite its sim-

[1] Two models are nonnested if one model cannot be reduced to the other model by imposing a set of linear restrictions on the parameter vector. See Clarke (2001) for an introduction to the issue of nonnested model testing.

AUTHOR'S NOTE: Earlier versions of this article were presented at the Workshop on Emerging Methodologies in the Quantitative Study of Conflict at Binghamton University (2001); the Peter D. Watson Center for Conflict and Cooperation Seminar Series at the University of Rochester (2001); and the International Relations Theory Colloquium at the University of California, Berkeley (2001). I thank the participants for their comments. I also thank Paul Huth, Todd Allee, and Brian Pollins for sharing their data and three anonymous reviewers for their comments. Errors remain my own.

JOURNAL OF CONFLICT RESOLUTION, Vol. 47 No. 1, February 2003 72-93
DOI: 10.1177/0022002702239512

plicity, outperforms the easiest and most common of the parametric tests under conditions that international relations scholars routinely face.

To demonstrate the usefulness of the method, I consider Monte Carlo evidence as well as two empirical debates in the international relations literature. The first is the debate over the effect of domestic politics on foreign policy decision making, and the second is the debate over long-cycle theory. (See the Two International Relations Examples section, below, for an introduction to these debates.) The results indicate that a political norms explanation outperforms a political accountability explanation and that long-cycle theory is insufficiently specified.

TWO RELATIVE DISCRIMINATION TESTS

There are three major approaches to testing nonnested models: a classical test of absolute discrimination, the Cox test; a classical test of relative discrimination, the Vuong test; and Bayes factors. With respect to the classical approaches, the difference between absolute and relative discrimination lies in the null hypothesis. Absolute discrimination tests, such as the Cox test, use one model as the null hypothesis and then use information from a rival model to test that null hypothesis. (Both models must serve as the null in turn, and both models may be rejected.) The null hypothesis for relative discrimination tests is that there is no significant difference between two models.

Of the three approaches noted above, the Vuong test is the most common, because it is the easiest of the three methods to compute beyond the realm of linear regression.[2] The Vuong test is also the least controversial of the three methods—the Cox test may reject both competing models (hypothesis tests generally decide between competing hypotheses), and Bayes factors depend on prior information. This study, then, focuses solely on the relative discrimination approach. In the next section, I demonstrate that the Vuong test does not perform well under conditions that international relations scholars routinely encounter.

THE VUONG APPROACH

The Vuong test is a relative discrimination test that is equivalent to a model selection test. Model selection procedures compare rival models directly by some criteria and then choose the "best" one. Model selection is unconcerned with whether the models perform well in absolute terms; what matters is the relative strength of the rival models.[3] That the Vuong test and the nonparametric test that follows in the next section are hypothesis tests represents an improvement over familiar model selection criteria such as R^2, C_p, Akaike's information criteria (AIC) (Akaike 1973), and Schwarz's

[2] Within the realm of linear regression, the linearized forms of the Cox test, the "J" and "JA" tests, are quite easy to implement (Doran 1993). An easy approximation for Bayes factors also exists, although its accuracy is questionable (Raftery 1995; Clarke 2000).

[3] The Vuong test, if it chooses a model, will choose the model that is closer to the true specification, even if both models are far from that specification.

(1978) Bayesian information criteria (BIC). Unlike the latter measures, the Vuong test and the nonparametric test provide probabilistic statements regarding model selection.

The Vuong test is based on the Kullback-Leibler (1951) information criteria (KLIC). Vuong defines the KLIC as

$$\text{KLIC} \equiv E_0[\ln h_0(Y_t|X_t)] - E_0[\ln f(Y_t|X_t;\ \beta_*)], \tag{1}$$

where $h_0(.|.)$ is the true conditional density of Y_t given X_t (that is, the true but unknown model), E_0 is the expectation under the true model, and β_* are the pseudo-true values of β (the estimates of β when $f[Y_t|X_t]$ is not the true model). The best model is the model that minimizes equation (1), for the best model is the one that is closest to the true specification. One should therefore choose the model that maximizes $E_0[\ln f(Y_t|X_t;\ \beta_*)]$. In other words, one model should be selected over another if the average log-likelihood of that model is significantly greater than the average log-likelihood of the rival model.

The null hypothesis of Vuong's test is

$$H_0 : E_0\left[\ln \frac{f(Y_t|X_t;\beta_*)}{g(Y_t|Z_t;\gamma_*)}\right] = 0, \tag{2}$$

meaning that the two models are equivalent.[4]

The expected value in the above hypothesis is unknown. Vuong demonstrates that under fairly general conditions,

$$\frac{1}{n} LR_n(\hat{\beta}_n, \hat{\gamma}_n) \overset{a.s.}{\to} E_0\left[\ln \frac{f(Y_t|X_t;\beta_*)}{g(Y_t|Z_t;\gamma_*)}\right], \tag{3}$$

which states that the expected value can be consistently estimated by $(\frac{1}{n})$ times the likelihood ratio statistic. The actual test is then

$$under\ H_0 : \frac{LR_n(\hat{\beta}_n, \hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} N(0,1), \tag{4}$$

where

$$LR_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv L_n^f(\hat{\beta}_n) - L_n^g(\hat{\gamma}_n) \tag{5}$$

and

$$\hat{\omega}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\ln \frac{f(Y_t|X_t;\hat{\beta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\ln \frac{f(Y_t|X_t;\hat{\beta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}\right]^2. \tag{6}$$

[4] γ_* and Z_t in model g are analogous to β_* and X_t in model f.

The Vuong test can be described in simple terms. If the null hypothesis is true, the average value of the log-likelihood ratio should be zero. If H_f is true, the average value of the log-likelihood ratio should be significantly greater than zero. If the reverse is true, the average value of the log-likelihood ratio should be significantly less than zero. In other words, the Vuong test statistic is simply the average log-likelihood ratio suitably normalized.

The log-likelihoods used in equation (4) can be affected if the number of coefficients in the two models being estimated is different, and therefore the test must be corrected for the degrees of freedom. Vuong (1989) suggests using a correction that corresponds to either Akaike's (1973) AIC or Schwarz's (1978) BIC.[5] I have chosen the latter, making the adjusted statistic[6]

$$L\tilde{R}_n(\hat{\beta}_n, \hat{\gamma}_n) \equiv LR_n(\hat{\beta}_n, \hat{\gamma}_n) - \left[\left(\frac{p}{2}\right)\ln n - \left(\frac{q}{2}\right)\ln n\right], \tag{7}$$

where p and q are the number of estimated coefficients in models f and g, respectively.

A Monte Carlo Experiment

How well the Vuong test performs under conditions actually encountered by international relations scholars is, of course, of interest. It is not uncommon for international relations studies to have small sample sizes (compared with voting studies) or for the competing models to be highly correlated (see Clarke 2001).[7]

A simple simulation was run to investigate the performance of the Vuong test under these conditions.[8] Two probit models, one of which is the data-generating process, are compared while controlling for the canonical correlation (ρ) between the models and the size of the sample. Canonical correlation is similar to bivariate correlation, except that instead of measuring the relationship between two variables, canonical correlation measures the relationship between two sets of variables (Johnson and Wichern 1998).

The results of the simulation are in Figure 1. For $n = 100$ and $\rho = .1$, the Vuong test chooses the true model less than 30% of the time. At $n = 200$, the probability rises to just over 70%; and at $n = 300$, just over 90%. Few, if any, competing models in the literature, however, have canonical correlations this low. As ρ increases, there is a steady

[5] The Akaike information criterion (AIC) and the Bayesian information criterion (BIC), like the adjusted R^2, are selection criteria that balance model fit with some adjustment for parsimony (Judge et al. 1985). These measures are particularly useful in a time-series framework where forecasting is the goal (Greene 1997). The problem with these measures is that one model will always be chosen even if neither model fits the data well (McAleer 1987). In addition, these techniques do not provide probabilistic statements regarding model selection. A further problem with using AIC is that it is generally not consistent (Schwarz 1978).

[6] Which correction factor is used makes no difference to this analysis.

[7] Seven recent small-n studies in conflict studies are Huth (1988), which has an n of 58; Huth, Gelpi, and Bennett (1993), which has an n of 97; Reiter and Stam (1998), which has an n of 197; Signorino and Tarar (2002), which has an n of 58; Bennett and Stam (1996), which has an n of 169; Benoit (1996), which has an n of 97; and Pollins (1996), which has an n of 161.

[8] An explanation of the simulation is given in the appendix.

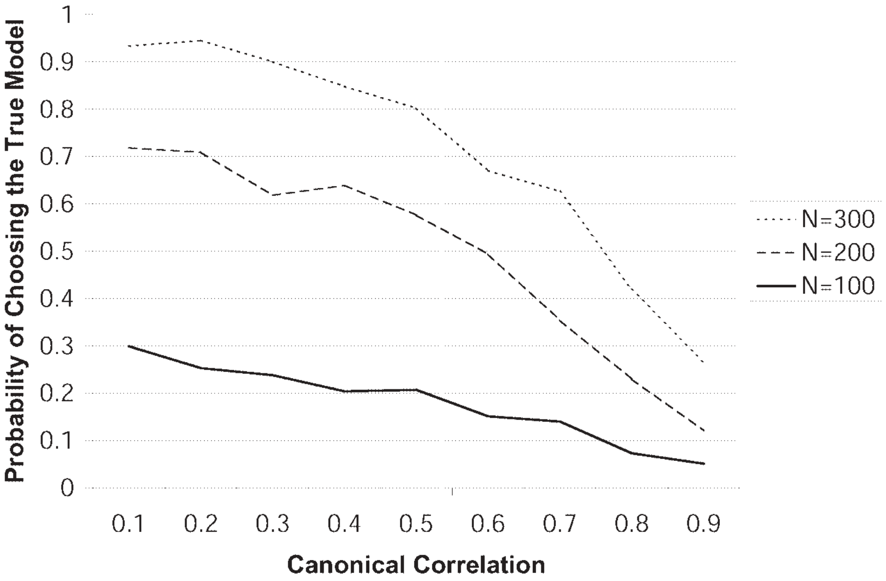

Figure 1: Canonical Correlation and the Vuong Test

diminution of the power of the test. At ρ = .6, the probability of the test choosing the true model drops to 15% for *n* = 100, 50% for *n* = 200, and less than 70% for *n* = 300. One might expect to see a decrease in the performance of any discrimination test as the models become more highly correlated (it is harder to distinguish between correlated models), but the Vuong test does not perform well even taking the effect of correlated models into account.

THE NONPARAMETRIC APPROACH

The Vuong test is not an exact test; it is only normally distributed asymptotically. The results of the experiment described above indicate that the relatively small sample sizes used in some international relations research may present a problem. One possible solution to this problem is to turn to nonparametric tests that do not rely on a priori parametric assumptions. Until very recently, few attempts have been made to apply nonparametric procedures to the problem of nonnested testing.[9] In what follows, I introduce a simple nonparametric test for nonnested models based on the paired sign test (Conover 1980).[10]

The Paired Sign Test

The paired sign test is one of the oldest and simplest of the nonparametric tests. It is used to test the null hypothesis that the probability of a random variable from the popu-

[9] Lavergne and Vuong (1996) is a notable exception.

[10] To my knowledge, this procedure is unknown and does not appear in the econometric or statistical literature.

lation of paired differences being greater than zero (or some other specific value) is equal to the probability of the random variable being less than zero. That is, the test is really a binomial test with $p = .5$. The test assumes only that the paired differences are independent.[11]

The paired test is similar to the better known Wilcoxon sign-rank test, which makes use of not just the signs of the differences but also their ranks (Wilcoxon 1945). The Wilcoxon test, however, makes the additional assumption that the distribution of paired differences is symmetric. When the symmetry assumption holds, the sign-rank test is the more powerful procedure. When the assumption does not hold and the paired differences follow a skewed distribution (such as the chi-square), the sign test is the more powerful test.[12]

A New Test for Relative Model Discrimination

The procedure I propose applies the paired sign test to the differences in the individual log-likelihoods from two nonnested models.[13] Whereas the Vuong test determines whether the average log-likelihood ratio is statistically different from zero, the proposed test determines whether the median log-likelihood ratio is statistically different from zero. If the models are equally close to the true specification, half the log-likelihood ratios should be greater than zero and half should be less than zero. If model *f* is "better" than model *g*, more than half the log-likelihood ratios should be greater than zero. Conversely, if model *g* is "better" than model *f*, more than half the log-likelihood ratios should be less than zero. The null hypothesis is therefore

$$H_0 : \theta = 0,$$

where θ is the median log-likelihood ratio.

One of the great strengths of this procedure is that implementing the test is remarkably simple and can be produced by any mainstream statistical software package using the following algorithm:[14]

1. Run model *f*, saving the individual log-likelihoods.
2. Run model *g*, saving the individual log-likelihoods.

[11] Note that this assumption does not imply that the samples are independent, only that the paired differences are independent. To guarantee consistency, one must also make the technical assumption that the paired differences all come from the same continuous distribution.

[12] For the test I propose, there is no reason to believe that this assumption holds, particularly in small samples. Simulations bear out this supposition.

[13] Recall that the log-likelihood reported by statistical software is the sum of the log-likelihoods for each individual observation. As a log-likelihood ratio is simply the difference between two log-likelihoods, there are *n* log-likelihood ratios—one for each observation.

[14] In what follows, steps 1 and 2 are accomplished easily by substituting predicted probabilities into the log-likelihood function. In the case of a binary choice model, the individual log-likelihoods are simply

$$\log L_i = y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i).$$

Steps 1 and 2 are in the process of being implemented in a future version of STATA. Steps 3 and 4 already exist in STATA, because I am making use of the paired sign test. The command is simply "signtest $ll_1 = ll_2$," where ll_i are the individual log-likelihoods from one model.

3. Compute the differences and count the number of positive and negative values.
4. The number of positive differences is distributed binomial (*n, p* = .5).

This test, like the Vuong test, may be affected if the number of coefficients in the two models being estimated is different. Once again, a correction for the degrees of freedom is needed. The Schwarz correction is

$$\left[\left(\frac{p}{2}\right)\ln n-\left(\frac{q}{2}\right)\ln n\right], \tag{8}$$

where *p* and *q* are the number of estimated coefficients in models *f* and *g*, respectively.

As I am working with the individual log-likelihood ratios, I cannot apply this correction to the “summed” log-likelihood ratio as I did for the Vuong test. I can, however, apply the average correction to the individual log-likelihood ratios. That is, I correct the individual log-likelihoods for model *f* by a factor of

$$\left(\frac{p}{2n}\right)\ln n,$$

and the individual log-likelihoods for model *g* by a factor of

$$\left(\frac{q}{2n}\right)\ln n.$$

In the Monte Carlo Results section, below, this simple test is seen to work exceedingly well. As with any nonparametric procedure, a loss of information (the magnitude of the differences) is traded for dubious parametric assumptions.

WHY NOT MAKE USE OF TRADITIONAL METHODS?

If one were to follow traditional practice in international relations scholarship, competing models would be tested by combining the rivals in a single equation.[15] The investigator would then use an *F* test or a likelihood ratio test to test a subset of variables. The problem with this approach can be demonstrated simply.[16]

Consider the following nonnested models with a common subset of variables:

$$H_f: \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_0 \tag{9}$$

[15] I address only a few of the more prevalent approaches due to space considerations. The arguments that follow have been made in greater depth elsewhere (see Clarke 2001), but it is useful to review them briefly here.

[16] This problem applies to both of the substantive examples I present later in the article.

$$H_g : \mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_1. \tag{10}$$

Combine these models into a single equation where $\tilde{\mathbf{X}}$ are the variables in **X** but not in **Z**, $\tilde{\mathbf{Z}}$ are the variables in **Z** but not in **X**, and **W** are the variables the two models have in common:

$$H_c : \mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\sigma} + \boldsymbol{\epsilon}.$$

Testing either $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ does not test the full models, because the variables in **W** are left out. On the other hand, testing $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ or $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ does not test H_g against H_f. The F test, in this case, discriminates between either H_f or H_g and a hybrid model that is neither H_f nor H_g (Kmenta 1986; Greene 1997).

Combining two separate theories in the same model is problematic even if the models do not have variables in common. The belief that rival models should be included in the same equation for “control” purposes is a bit of folklore that has no basis in statistical theory. None of the econometric texts commonly used by political scientists (Greene 1997; Davidson and MacKinnon 1981; Kmenta 1986; Judge et al. 1985) recommends this approach. Theory should dictate the choice of covariates in a model, and including the variables from both models in the same equation amounts to misspecification. Unless there exists a well-specified theory that claims that both models are relevant to the phenomena in question, the two sets of covariates do not belong in the same equation.[17]

Another approach to model discrimination, as noted earlier, is to make use of model selection criteria such as the AIC and BIC. The problem with these measures is that neither provides probabilistic statements regarding model selection. Whether the difference between the AIC values for two models is large is unknown, because there is no distribution associated with the measure. A further problem with these measures is that unlike the nonnested tests, they do not include information from the rival model.

MONTE CARLO RESULTS

One can reasonably expect that the performance of any discrimination procedure may be affected by the size of the sample, the correlation between the competing models, and the dimensions of the competing models. How the tests perform under these conditions is of great interest. To answer these questions, I performed simulations to gauge the effects of sample size, canonical correlation, and relative model dimensions on the two tests.[18]

[17] If we take the Pollins (1996) models in the Two International Relations Examples section, below, as an example, the models could not be combined even if doing so were legitimate, because the combined matrix of covariates does not have full column rank.

[18] Additional simulations were performed to assess the effects of autocorrelation and nonnormal disturbances. The results are available on request—neither violation had a significant effect on the tests under consideration.

Although other approaches to nonnested discrimination, such as the Cox test and Bayes factors, have undergone extensive testing, the Vuong test does not appear to have been the subject of any Monte Carlo tests (despite a call for such tests by Vuong himself [Vuong 1989]). The results that follow, therefore, cannot be compared with any previously published results.

THE FRAMEWORK OF THE SIMULATION

The experiments are based on the discrimination of two simple probit models, one of which is the data-generating process. The probit was chosen for convenience and because of its ubiquity in international relations research. Details of the simulation are given in the appendix.

When performing Monte Carlo experiments, one is generally interested in the size of the test, the probability of rejecting a true null, and power of the test, the probability of rejecting a false null (Mooney 1997). Power is well defined for both the Vuong and nonparametric tests. The null hypothesis is false in every experiment, and one can easily calculate the probability of rejecting the false null. Size, however, is not well defined. With the design given in the appendix, the null hypothesis is never true. That is, the models are never actually equal.[19] The concept of test size, the probability of rejecting a true null, therefore has no meaning.

A framework for comparing these tests emerges when one considers the probabilities that are produced by the simulations. Both tests produce three probabilities, listed below, which can be used for comparison in place of the traditional probabilities:

1. the probability of rejecting H_0 in favor of the true model,
2. the probability of rejecting H_0 in favor of the false model,
3. the probability of failing to reject H_0.

I can now talk of the probability of choosing the correct model, the probability of choosing the wrong model, and the probability of choosing neither model.

RESULTS OF THE SIMULATION

Canonical Correlation

The first simulation concerns the effect of the degree of correlation across the competing models. The null and alternative are the base models given in the appendix. Fifty-four experiments were performed with the following parameters:

$n = (100, 200, 300)$
$\rho_1^* = (.1, .2, .3, .4, .5, .6, .7, .8, .9)$

[19] A future experiment may include a situation where the data-generating process is a combination of two models: $y^* = \alpha(X\beta) + (1 - \alpha)(Z\gamma) + \varepsilon$.

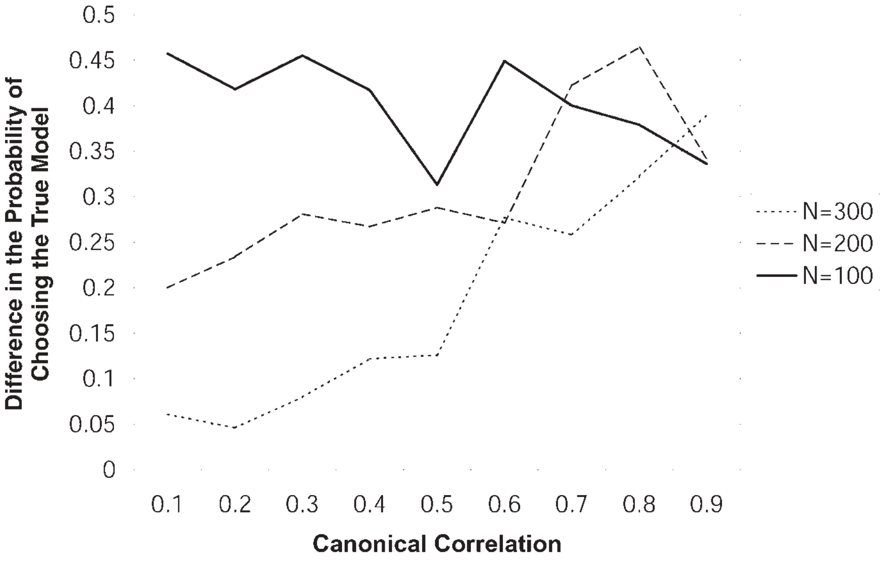

Figure 2: Canonical Correlation Comparison

$\sigma = (4)$
Methods = (Vuong, nonparametric)

Each experiment was replicated 1,000 times.[20]

The results for $n = 100, 200$, and 300 are in Figure 2. Note that Figure 2 gives the *difference* (nonparametric minus Vuong) for each sample size in the probability of choosing the true model for the two tests. If the tests performed equally well, each series would be at zero.[21] The nonparametric test, however, outperforms the Vuong test at every level of correlation. As the correlation increases, the difference in the performance of the tests increases dramatically at the $n = 100$ and 200 levels. (At $n = 300$, the difference hovers around .4.)

As expected, the difference between the tests narrows as the sample size increases. Even at $n = 300$, however, the Vuong test chooses the correct model less than 80% of the time as the correlation rises above .5. At the extreme correlation of .9, the nonparametric test is choosing the true model more than twice as often as the Vuong test (.66 v. .27).

As just demonstrated, the nonparametric tests chooses the true model more often than the Vuong test. It also chooses the false model more often than the Vuong test. The probability of the tests choosing the false model are given in Table 1. As expected, the

[20] These numbers can be justified by noting that if α is the probability of a type I error and $\hat{\alpha}$ is its estimate obtained from a simulation with v replications, the 95% confidence interval, $\left(|\alpha - \hat{\alpha}|/\sqrt{[\alpha(1-\alpha)]/v}\right) = 1.96$, gives $v = 456$ for $\alpha = .05$ and $\alpha - \hat{\alpha} = \pm.02$, and $v = 811$ for $\alpha = .05$ and $\alpha - \hat{\alpha} = \pm.015$.

[21] This is true of Figure 3 as well.

TABLE 1
The Probability of Choosing the False Model

r	n	*Vuong*	*Nonparametric*
.4	100	0	.009
	200	0	0
	300	0	0
.5	100	0	.002
	200	0	0
	300	0	.001
.6	100	0	.009
	200	0	.001
	300	0	0
.7	100	0	.015
	200	0	.001
	300	0	.001
.8	100	.001	.023
	200	0	.008
	300	0	.005
.9	100	.002	.028
	200	0	.02
	300	0	.007

probability is affected by the level of correlation between the models, the size of the sample, and the size of the error term variance.

The Vuong test rarely chooses the false model even under the most adverse conditions (high canonical correlation and small sample size). Neither method, however, often chooses the false model in absolute terms. The fact that the Vuong test does relatively better than the nonparametric test in not choosing the false model does not offset its worse performance in choosing the true model.

Models with Different Dimensions

Nonnested tests, as I noted, can be affected by the relative dimensionality of the rival models. Correction factors are built into the Vuong and nonparametric tests (both use the BIC), but how well these correction factors work is still in question.

Three sets of models were created: the first with four variables in the true model and two in the false model, the second with four variables in both models, and the third with four variables in the true model and six in the false model. In each case, the canonical correlation between the models is set at .5.[22]

For this experiment, the data-generating process given in equation (14) (in the appendix) is altered to include four variables:

[22] Creating rival models with different dimensions while controlling the canonical correlation calls for a nonsquare matrix, ρ_{12}. See the appendix.

$$y^* = .2 - 1.9x_1 + .25x_2 + .3x_3 + .1x_4 + \varepsilon_0,\ \varepsilon_0 \sim N(0, 4), \tag{11}$$

where y^* is unobserved as before.

Eighteen experiments were performed with the following parameters:

n = (100, 200, 300)
$\rho_1^* = (.5)$
$\sigma = (4)$
Model dimensions: 4 × 2, 4 × 4, 4 × 6
Methods = (Vuong, nonparametric)

Each experiment was replicated 1,000 times.

The results for the relative discrimination methods for the three sample sizes are in Figure 3. Again, the nonparametric test outperforms the Vuong test, and as expected, the difference in the probability of choosing the true model decreases as the sample size increases. The difference in the probability of choosing the true model also decreases as the relative dimensions of the models move in favor of the false model and away from the true model. The difference must decrease as the probability of choosing the true model for both tests approaches 1. The explanation lies in the BIC correction factor shared by the two methods.

In the case where the false model has more parameters than the true model, the correction factor works in favor of the true model. The probability of choosing the true model is greatest in this case because the false model is penalized for additional irrelevant variables. The correction factor is so strong that the tests work better in this situation than when the models have equal parameters.

The flip side, of course, is when the true model has more parameters than the false model. In this situation, the strength of the correction factor works against the tests. The probability of choosing the true model decreases from the case where the parameters are equal to the case where the true model has more parameters. Again, this effect is ameliorated as the sample size increases and would disappear were the sample size large enough.

DISCUSSION

The major finding here is that the nonparametric test significantly outperforms the Vuong test. This statement is true across all the experiments, regardless of the sample sizes used. Highly correlated models, as expected, are harder to distinguish than more independent models; and the relative dimensions of the competing models do have an effect on the performance of the tests. (This last effect is mitigated somewhat by use of the BIC correction factor.)

The difference between the tests, however, would disappear quickly for sample sizes greater than 300. This fact reflects the asymptotic nature of the Vuong test. Small sample sizes, however, do crop up in international relations research, and for the purposes of such research, the Vuong test has little to recommend it over the

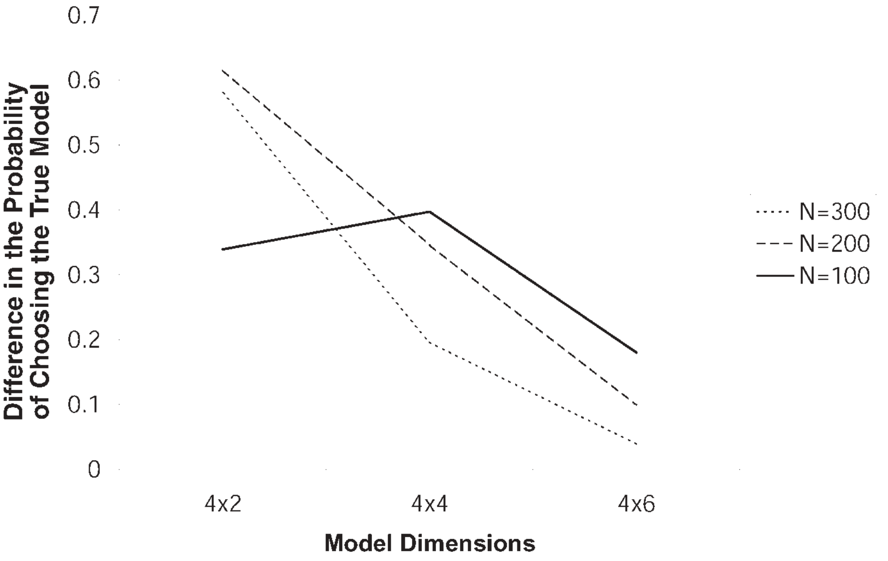

Figure 3: Model Dimension Comparison

nonparametric test. The single strength of the Vuong test is that it rarely chooses the false model. As noted earlier, this virtue does not outweigh the poor performance of the test under these conditions. As highly correlated models are common in international relations research, the nonparametric test should be used whenever the sample size is smaller than 400.

TWO INTERNATIONAL RELATIONS EXAMPLES

DOMESTIC POLITICS AND FOREIGN POLICY DECISION MAKING

These data are taken from a forthcoming book by Paul Huth and Todd Allee.[23] The broad object of the project is to determine the effect of domestic institutions on foreign policy decision making. Huth and Allee compare three models, each of which corresponds to a different causal mechanism that links domestic institutions to foreign policy decisions. In the first, the political accountability model, institutions are a source of political accountability for decision makers. In the second, the political norms model, institutions are a source of norms for bargaining in international conflict. In the third model, the political affinity model, the similarity of institutions between states is a

[23] I am using these data solely for illustrative purposes. My analysis was performed before Huth and Allee completed their analysis. My specifications, therefore, differ considerably from the more theoretically informed specifications used in their forthcoming book. These differences have no doubt affected some inferences.

source of international threat perception. I focus only on the political accountability and political norms models.

In the political accountability model, competitive elections, independent legislative powers, and the threat of military coups are the sources of accountability for leadership decisions in foreign policy. Huth and Allee (forthcoming) base the political accountability model on four key assumptions: (1) the critical goal of incumbent leaders is the retention of their office; (2) political opponents challenge incumbents at strategic junctures; (3) political accountability varies across different domestic political institutions, and (4) the greater the political vulnerability of leaders, the more risk-averse leaders are in their foreign policy.

I operationalize the political accountability model using five variables: a dummy variable for whether the state is democratic, a variable for how secure the governing party is,[24] a dummy variable for whether the leader was in a close election within 6 months,[25] and an interaction term for whether the state has experienced a coup within the past 2 years and is nondemocratic.

In the political norms model, "attention shifts to the principles that shape political elite beliefs about how to bargain and resolve political conflicts," and leaders from democratic and nondemocratic states have "different beliefs about the acceptability of compromising with and coercing political adversaries" (Huth and Allee forthcoming). Huth and Allee (forthcoming) base the political norms model on three main assumptions: (1) norms influence decisions made by political actors in political conflict, (2) domestic political institutions structure political conflict, and (3) the bargaining strategies used by leaders in international disputes are influenced by the norms of bargaining those same leaders use with domestic political opponents.[26]

I operationalize the political norms model using three variables: the strength of democratic norms in the state,[27] a dummy variable for whether the state has recently become democratic (within the past 5 years), and a dummy variable for whether the state is very nondemocratic.[28]

These competing models are matched with a straightforward realist model comprising military balance, alliance behavior, strategic interests, and previous dispute behavior. The combined models are tested on territorial disputes where the challenger has opted for military pressure over calling for negotiations. The challenger and the target both choose to either escalate the dispute or not (the dependent variable). The models are consequently estimated with a bivariate probit. The results for the political accountability model and the political norms model are in Tables 2 and 3, respectively.

[24] The percentage of seats held by the governing coalition in the lower house.

[25] "Close" means less than 5% of the vote.

[26] "Norms" is used by Huth and Allee (forthcoming) to refer to "principles or standards concerning which political actions and behaviors are seen as legitimate and desirable when engaging in political competition and seeking to resolve political conflict."

[27] Number of the past 20 years the state has been democratic.

[28] A Polity 98 netdemocracy score of –5 or less for at least 10 of the past 20 years.

TABLE 2
The Political Accountability Model (n = 389)

	Challenger		*Target*	
Variable	*Coefficient*	SE	*Coefficient*	SE
Realist variables				
Alliance	–.404	.191	–.099	.194
Strategic value	.391	.185	–.105	.166
Military ratio	.363	.276	–.371	.284
Other conflict	.176	.143	.235	.155
Previous stalemate	–.306	.148	–.142	.149
Domestic variables				
Democratic	–.26	.507	–.746	.413
Security of leaders	.005	.007	.013	.006
Recent close race	4.95	0	–.547	.672
Nondemocratic × No Coup	.088	.158	.075	.177
Constant	.581	.193	.886	.239
ρ	.77	.053		
Log-likelihood	–351.6			

TABLE 3
The Political Norms Model (n = 389)

	Challenger		*Target*	
Variable	*Coefficient*	SE	*Coefficient*	SE
Realist variables				
Alliance	–.421	.187	–.183	.192
Strategic value	.516	.184	–.046	.167
Military ratio	.352	.281	–.563	.297
Other conflict	.166	.144	.274	.155
Previous stalemate	–.322	.148	–.099	.148
Domestic variables				
Strength of norms	–.03	.018	–.024	.012
Recently democratic	.045	.261	–.111	.302
Most Nondemocratic	–.446	.161	–.235	.178
Constant	.944	.207	1.250	.252
ρ	.77	.052		
Log-likelihood	–348.5			

Results

Reporting the results from a discrimination test differs substantially from reporting the results from a regression, for example. In the latter situation, a discussion of the substantive effects of specific variables should accompany the statistical results. Discrimination tests, however, are generally applied after a researcher has demonstrated that each of the competing models has a correspondence with the data. Discussion of the effects of specific variables, therefore, is unnecessary, because such a discussion has generally already taken place. Furthermore, application of a discrimination test does not change the results of the individual analyses. Substantive discussion of the results for the individual models, therefore, is beyond the scope of this study, and I refer the reader to Huth and Allee (forthcoming) and Pollins (1996).

The results of the Vuong discrimination test for the Huth and Allee models are in Table 4. The test returns a statistic of –.18 and a confidence interval comfortably bracketing zero. I therefore fail to reject the null hypothesis of "no difference" at conventional significance levels and conclude that the models explain equally well. Given the canonical correlation between the models ($\rho = .85$), it is not surprising that the Vuong test cannot distinguish between them.

The results of the nonparametric test for the Huth and Allee models are in Table 5. Where the Vuong test could not distinguish between these highly correlated models, the nonparametric test readily distinguishes between them. I reject the null hypothesis of "no difference" in favor of the political norms model at conventional significance levels. The fact that the nonparametric procedure can distinguish between these models, given the complex nature of the estimation, the demands bivariate probit makes on the data, and the relatively small sample size, makes a strong statement about the utility of the procedure. In addition, the finding that a political norms explanation has greater explanatory power than a political structure explanation corroborates earlier work on the effect of domestic politics on foreign policy decision making (see Maoz and Russett 1993).

LONG CYCLES IN INTERNATIONAL RELATIONS

Long cycles have been a key explanatory mechanism for scholars arguing for systemic explanations of great power war.[29] There has never been, however, much agreement on which long cycles are important or how to measure them. Two broad approaches have been taken to the issue. The first focuses on global economic activity and is championed by Joshua Goldstein (1991). The second focuses on the global political order and characterizes work by Immanuel Wallerstein (1983), Modelski and Thompson (1987), and Robert Gilpin (1981). As each model has received significant empirical corroboration (Pollins 1996), the question of which model best explains great power war is of significant interest. Pollins (1996), however, takes issue with both approaches and argues that the focus on systemic war should be abandoned in

[29] See Pollins (1996) for a large number of cites available on this topic.

TABLE 4
Results of the Vuong Test for the Huth Models

Vuong	SE	Z *Statistic*	*Significance*	*95% Confidence Interval*
–0.180	0.24	–0.747	.455	–0.65 to 0.29

TABLE 5
Results of the Nonparametric Test for the Huth Models

One-sided tests[a]
 H_1: median of model $f - g > 0$
  Binomial($n = 389$, $x \geq 174$, $p = .5$) = .9835
 H_1: median of model $f - g < 0$
  Binomial($n = 389$, $x \geq 215$, $p = .5$) = .0212
Two-sided test
 H_1: median of model $f - g \neq 0$
  Min[1, 2 × Binomial($n = 389$, $x \geq 215$, $p = .5$)] = .0424

a. H_0: median of $f - g = 0$ for all tests.

favor of including lower level conflict and that both the economic and political long cycles should be modeled as interlinked coevolving systems.

To test his theory against what he refers to as the “core” frameworks, Pollins (1996) develops statistical models of each of the five theories noted above. I focus only on the Goldstein (1991) model and the coevolving systems model; the empirical results reported by Pollins for these models are in Table 6. Goldstein’s model is based on his contention that periods of economic upswing coincide with major power conflict and that periods of economic downswing are relatively peaceful. Goldstein unpacks the long wave into four phases labeled expansion, war, stagnation, and rebirth. The first two phases correspond to upswings in economic activity, and the latter two correspond to downswings. Pollins models Goldstein’s theory simply by using dummy variables for each of Goldstein’s four phases. Pollins also includes, as control variables, the number of states in the system, as well as the dependent variable lagged, to address the time dependence inherent in the models (p. 110).

Pollins’s (1996) model of evolving systems is based on his argument that the level of armed conflict is influenced by a combination of economic and political long cycles. In particular, the model integrates the economic long cycles determined by Goldstein (1991) with the political long cycles determined by Modelski and Thompson (1987).[30] Systemic conflict will be lowest when each cycle is at its most peaceful point—stagna-

[30] This is not to imply that the Goldstein (1991) model is nested within the coevolving systems model. This point will be made clear in the model description.

TABLE 6
Poisson Models of Systemic Conflict (n = 161)

	Model 1		*Model 2*	
Variable	*Coefficient*	SE	*Coefficient*	SE
Goldstein				
Stagnation	0.645*	0.1		
Rebirth	0.857*	0.096		
Expansion	1.068*	0.085		
War	1.100*	0.105		
Lagged dependent	0.061*	0.009		
Members	0.006*	0.001		
Coevolving systems				
CR2			–0.023	0.238
CR3			0.557*	0.119
CR4			0.939*	0.109
CR5			0.749*	0.111
CR6			1.198*	0.094
CR7			0.853*	0.122
CR8			1.464*	0.141
Lagged dependent			0.044*	0.01
Members			0.009*	0.002
Log-likelihood	–368.6		–348.3	

NOTE: CR*n* is a dummy variable that takes a one when the year has a cumulative rank ordering equal to *n*.
*p < .001.

tion for Goldstein and world power for Modelski and Thompson. Because these two frameworks sometimes conflict for a given year (1855 is an example), Pollins rank-orders the level of armed conflict in the four periods specified by each framework and then sums these rank-orderings. Each year, then, is coded (from 2 to 8) according to the cumulative effect that the two cycles claim to have.[31] Pollins models his theory by including seven dummy variables, one for each possible rank-ordering. The same control variables used in the Goldstein equation are also included.

Each model appears to fit the data quite well. In a pairwise likelihood ratio test, Pollins (1996) concludes that his coevolving systems model is superior to Goldstein's (1991) model. The models, however, are clearly nonnested in their covariates, and the likelihood ratio test may only be applied to nested models.

[31] Pollins (1996) provides the following examples:

> The year 1915, for example, is coded 8 because it falls within the periods of highest expected conflict within both frameworks. Similarly, 1820 receives a score of 2 because this year falls with periods of lowest conflict with regard to both the economic long wave and the world leadership cycle. (P. 111)

TABLE 7
Results of the Vuong Test for the Pollins Models

Vuong	SE	Z *Statistic*	*Significance*	*95% Confidence Interval*
–12.6	7.58	–1.67	.1	–27.46 to 2.26

TABLE 8
Results of the Nonparametric Test for the Pollins Models

One-sided tests[a]
H_1: median of model $f - g > 0$
Binomial($n = 161$, $x \geq 86$, $p = .5$) = .22
H_1: median of model $f - g < 0$
Binomial($n = 161$, $x \geq 75$, $p = .5$) = .83
Two-sided test
H_1: median of model $f - g \neq 0$
Min[1, 2 × Binomial($n = 161$, $x \geq 86$, $p = .5$)] = .43

a. H_0: median of $f - g = 0$ for all tests.

Results

The results of the Vuong test for Pollins's (1996) models are in Table 7. The test returns a statistic of –12.6, which appears to support Pollins's contention that the coevolving systems model outperforms its rival. The 95% confidence interval does, however, include zero, and therefore the null hypothesis cannot be rejected technically at any significance level below 10%. With reservations then, the inference Pollins made can be accepted.

The nonparametric results for the Pollins (1996) models (Table 8) echo the results of the Vuong test in the sense that the null hypothesis of equality cannot be rejected in either direction. The *p* value for the nonparametric test, however, is larger than the *p* value for the Vuong test. Furthermore, unlike the Vuong results, the test results appear to lean toward the Goldstein (1991) model. These findings are odd given the results of the Monte Carlo experiments. One would expect the nonparametric test to do a better job of discriminating between the models. The anomalous results are explained easily when one considers the difference in the way these tests treat outliers. Looking at the individual log-likelihood ratios generated by the tests reveals three negative values that are more than four standard deviations away from the mean of the series, –.0785. These values are affecting the Vuong test, which is based on the mean, far more than they affect the nonparametric test, which is based on the median. In effect, these values are dragging down the mean of the log-likelihood ratio and thereby causing the large negative Vuong statistic. The robustness of the nonparametric method in the face of such outliers is an added bonus of the procedure.

CONCLUSION

International relations scholars face a number of problems when trying to discriminate between rival models. These problems include highly correlated models and often small sample sizes. The common parametric approach to relative model discrimination, the Vuong test, does not perform well under these conditions. In this study, I proposed a simple nonparametric test for relative model discrimination in place of the Vuong test. The nonparametric test has a number of specific advantages over the Vuong test. First, Monte Carlo experiments demonstrate that the nonparametric test performs far better than the Vuong test under conditions, such as small sample sizes and high model correlation, encountered by international relations scholars. Second, the procedure is trivial to implement even for the most complicated models. Third, the nonparametric test is more robust than the Vuong test in the presence of outliers. Two empirical examples concerning long cycles and the effect of domestic politics on foreign policy decision making demonstrate the points made above. In both examples, the nonparametric test provides a clearer picture than the Vuong test of the relative merits of the competing models.

APPENDIX
Design of the Simulation

SETTING THE CANONICAL CORRELATION

The procedure for setting the canonical correlation between two sets of variables **X** and **Z** is to multiply orthogonal random variates by the Cholesky decomposition of a correlation matrix, ρ (Kaiser and Dickman 1962). Let $\mathbf{X} = (x_1, x_2)$ and $\mathbf{Z} = (z_1, z_2)$ be two sets of variables, and let $\mathbf{C} = (c_1, c_2, c_3, c_4)$ be four standardized orthogonal variables. Let

$$\rho = \begin{bmatrix} \rho_{11} & \vdots & \rho_{12} \\ \cdots & \cdot & \cdots \\ \rho_{21} & \vdots & \rho_{22} \end{bmatrix}$$

be the overall correlation matrix where ρ_{ij} is a 2 × 2 correlation matrix. If **A** is the Cholesky decomposition of ρ, then **X** and **Z** are produced by the following relation:

$$\left[\mathbf{X} \vdots \mathbf{Z}\right] = \mathbf{CA}. \qquad (12)$$

A common method of solving for ρ_{12} is to use the following equation given by Johnson and Wichern (1998):

$$\left|\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1}\rho_{21} - \lambda_1 I\right| = 0\,, \qquad (13)$$

where λ_1, the largest eigenvalue, is equal to the square of the first canonical correlation, $\lambda_1 = \rho_1^{*2}$. For a canonical correlation of .5, set $\lambda_1 = .25$, let $\rho_{12} = [\begin{smallmatrix} x & 0 \\ 0 & 0 \end{smallmatrix}]$, and solve equation (13) for x. Note that ρ_{12} need not be a square matrix. Creating an **X** with four columns and a **Z** with two columns with canonical correlation, ρ_1^*, simply requires expanding ρ_{11}, ρ_{12}, and λI appropriately.

THE MODELS

The null and alternative models each contain two normally distributed variables, $\mathbf{X} = (x_1, x_2)$ and $\mathbf{Z} = (z_1, z_2)$, that are correlated at .4 and .2, respectively.[32] The canonical correlation is set by the procedure in the previous section.

The data were generated according to the following model:

$$y^* = .2 - 1.9x_1 + .25x_2 + \varepsilon_0,\ \varepsilon_0 \sim N(0, 4), \tag{14}$$

where y^* is unobserved:

$$y = \begin{cases} 1 \text{ if } y^* > 0 \\ 0 \text{ if } y^* \leq 0 \end{cases}.$$

The null hypothesis for each test is that there is no difference in the "closeness" of the models to the data-generating process. The null may be rejected in favor of either model. Equation (14) is modified for the experiment concerning dimensionality, and these changes are noted in the appropriate section.

REFERENCES

Akaike, H. 1973. Information theory and an extension of the likelihood ratio principle. In *Second international symposium of information theory* (Minnesota Studies in the Philosophy of Science), edited by B. N. Petrov and F. Csaki, 267-81. Budapest, Hungary: Akademinai Kiado.

Bennett, D. Scott, and Allan C. Stam. 1996. The duration of interstate wars, 1816-1985. *American Political Science Review* 90:239-57.

Benoit, Kenneth. 1996. Democracies really are more pacific (in general): Reexamining regime type and war involvement. *Journal of Conflict Resolution* 40:636-57.

Clarke, Kevin A. 2000. The effect of priors on approximate Bayes factors from mcmc output. Unpublished manuscript.

———. 2001. Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science* 45:724-44.

Conover, W. J. 1980. *Practical nonparametric statistics*. 2d ed. New York: John Wiley.

Davidson, Russell, and James G. MacKinnon. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49:781-93.

Doran, Howard. 1993. Testing nonnested models. *American Journal of Agricultural Economics* 75:95-103.

Gilpin, Robert. 1981. *War and change in world politics*. Cambridge: Cambridge University Press.

Goldstein, Joshua. 1991. A war-economy theory of the long wave. In *Business cycles: Theories, evidence, and analysis*, edited by N. Thygesen, K. Velupillai, and S. Zambelli, chap. 12. London: Macmillan.

Greene, William H. 1997. *Econometric analysis*. 3d ed. Englewood Cliffs, NJ: Prentice Hall.

Huth, Paul K. 1988. *Extended deterrence and the prevention of war.* New Haven, CT: Yale University Press.

Huth, Paul K., and Todd Allee. Forthcoming. *The democratic peace and territorial conflict in the twentieth century.*

[32] As long as these correlations are set at levels below which multicollinearity is likely, their values do not matter.

Huth, Paul, Christopher Gelpi, and D. Scott Bennett. 1993. The escalation of great power militarized disputes: Testing rational deterrence theory and structural realism. *American Political Science Review* 87:609-23.

Johnson, Richard A., and Dean W. Wichern. 1998. *Applied multivariate statistical analysis*. 4th ed. Englewood Cliffs, NJ: Prentice Hall.

Judge, George G., W. E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee. 1985. *The theory and practice of econometrics*. 2d ed. New York: John Wiley.

Kaiser, Henry F., and Kern Dickman. 1962. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika* 27:179-82.

Kmenta, Jan. 1986. *Elements of econometrics*. 2d ed. New York: Macmillan.

Kullback, Solomon, and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22:79-86.

Lavergne, Pascal, and Quang H. Vuong. 1996. Nonparametric selection of regressors: The nonnested case. *Econometrica* 64:207-19.

Maoz, Zeev, and Bruce Russett. 1993. Normative and structural causes of democratic peace, 1946-1986. *American Political Science Review* 87:624-38.

McAleer, Michael. 1987. Specification tests for separate models: A survey. In *Specification analysis in the linear model*, edited by M. L. King and D. E. A. Giles. London: Routledge and Kegan Paul.

Modelski, George, and William R. Thompson. 1987. Testing cobweb models of the long cycle. In *Exploring long cycles*, edited by George Modelski, 85-111. Boulder, CO: Lynn Rienner.

Mooney, Christopher Z. 1997. *Monte Carlo simulation: Quantitative application in the social sciences*. Thousand Oaks, CA: Sage.

Pollins, Brian M. 1996. Global political order, economic change, and armed conflict: Coevolving systems and the use of force. *American Political Science Review* 90:103-17.

Raftery, Adrian E. 1995. Bayesian model selection in social research (with discussion). In *Sociological methodology 1995*, edited by P. V. Marsden. Cambridge, MA: Blackwell.

Reiter, Dan, and Allan Stam. 1998. Democracy, war initiation, and victory. *American Political Science Review* 92:377-89.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6:461-64.

Signorino, Curtis S., and Ahmer Tarar. 2002. A unified theory and test of extended immediate deterrence. Unpublished manuscript.

Vuong, Quang. 1989. Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* 57:307-33.

Wallerstein, Immanuel. 1983. Three instances of hegemony in the history of the capitalist-world economy. *International Journal of Comparative Sociology* 24:100-108.

Wilcoxon, Fred. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1:80-83.