Testing Nonnested Models of International Relations: Reevaluating Realism

Kevin A. Clarke University of Michigan

Unknown to most world politics scholars and political scientists in general, traditional methods of model discrimination such as likelihood ratio tests, F-tests, and artificial nesting fail when applied to nonnested models. That the vast majority of models used throughout international relations research have nonlinear functional forms complicates the problem. The purpose of this research is to suggest methods of properly discriminating between nonnested models and then to demonstrate how these techniques can shed light on substantive debates in international relations. Reanalysis of two well-known articles that compare structural realism to various alternatives suggests that the evidence against realism in both articles is overstated.

Is the escalation of great-power militarized disputes better explained by rational deterrence theory or by structural realism? Are war outcomes better explained by regime type and initiation effects or by traditional realist variables? Each of these debates has been the subject of recent articles in a leading political science journal (Huth, Gelpi, and Bennett 1993; Reiter and Stam 1998b). The conclusions reported in these articles have added to the growing presumption that realism is wrong or simply insufficient to account for conflict outcomes in international relations. At the heart of each article is a statistical comparison of rival models. The techniques used by these authors, however, cannot perform the required comparisons because the models being compared are nonnested or "separate." Two models are nonnested if one model is not a special case of the other model (see the third section for a precise definition).

Unknown to most world politics scholars and political scientists in general, traditional methods of model discrimination such as likelihood-ratio tests, F-tests, or artificial nesting fail when applied to nonnested models. That the vast majority of models used throughout international relations are nonlinear in terms of their functional forms only complicates the situation. The purpose of this research is to suggest techniques that discriminate properly between nonnested models and then demonstrate how these techniques can shed light on the aforementioned debates. The results of my analyses suggest that the evidence against realism in both articles is overstated.

Kevin A. Clarke is a Ph.D. candidate in Political Science, University of Michigan, 611 Church Street, Ste. 334, Ann Arbor, MI 48104-3028 (kclarke@umich.edu).

Earlier versions of this paper have been presented at the annual meetings of the American Political Science Association in Boston, MA (1998) and in Atlanta, GA (1999), and at the annual meetings of the Society for Political Methodology in San Diego, CA (1998) and in College Station, TX (1999); I thank the participants for their comments. I also thank John Jackson, Paul Huth, Bob Pahre, Susan Murphy, Laura Koehly, Ed Czilli, and several anonymous referees for helpful comments and discussions. Errors remain my own. Paul Huth, Christopher Gelpi, D. Scott Bennett, Dan Reiter, and Al Stam generously shared their data. Heather Edes was gracious enough to read numerous drafts. Nick Winter provided invaluable programming assistance and Henry Heitowit provided encouragement and a quiet place to work.

American Journal of Political Science, Vol. 45, No. 3, July 2001, Pp. 724–744

Two Debates in International Relations

The two articles upon which I have chosen to concentrate no longer represent the state of the art in their respective debates. The articles have been surpassed by research that explicitly models the strategic interactions inherent in the theories. Smith (1999) and Signorino (1999) show that analyzing strategic theories with traditional statistical models results in misspecification.[1] The Smith and Signorino papers demonstrate that statistical models that take strategic interaction into account outperform statistical models that do not. I am putting aside the strategic issues, however, in the interest of a clean exposition and simpler mathematics. My goal is to suggest corrections to a common problem in the empirical literature of international relations, not to correct theoretical problems.

Structural Realism versus Rational Deterrence Theory

Does structural realism or rational deterrence theory provide a better explanation of the escalation of great-power militarized disputes? Realism has been the dominant theoretical position in international relations for the last fifty years, and structural realism (Waltz 1979) has been the dominant brand of realism for the past twenty years. Rational deterrence theory (Schelling 1960) has been a serious contender for that position. Properly specified, both theories demonstrate an empirical grasp on important problems in world politics (Huth, Bennett, and Gelpi 1992). Efforts to test these theories against one another, however, have ignored the fact that the theories are nonnested, and therefore standard model selection techniques are inappropriate.

The most systematic attempt to test structural realism against rational deterrence theory can be found in Huth, Bennett, and Gelpi (1993). In that article, Huth and his co-authors conceptualize structural realism in terms of the amount of uncertainty created by the structure of the international system. To connect the amount of uncertainty in the system to actual decisions taken by state leaders, the authors interact uncertainty with the risk propensities of these decision makers. When uncertainty is high, risk-acceptant leaders will pursue policies that might spark armed conflict, while risk-averse leaders will likely be more cautious (Huth, Bennett, and Gelpi 1993). Structural realism, then, is operationalized by two composite measures of uncertainty (size and capability diffusion), a measure of risk propensity, and two interaction terms (one for each measure of uncertainty).[2]

As for rational deterrence theory, Huth, Gelpi, and Bennett (1993, 612) argue that, "the credibility of the threat is the primary determinant of deterrence success or failure." Credibility is affected by the balance of military capabilities, the interests at stake for the states involved, the past dispute behavior of the states, and whether either state is engaged in another dispute at the same time. Deterrence is more likely to fail as the balance of capabilities and the interests at stake shift toward the challenger. Deterrence is also more likely to fail if the defender has backed down in a previous dispute or is engaged in a dispute elsewhere.

Huth, Gelpi, and Bennett (1993, 619) compare their models by combining them in a single, large equation. The authors conclude that "rational deterrence theory provides a much more compelling explanation of great-power decisions to escalate militarized disputes than does structural realism." A replication of their results is in Table 1. Their conclusion is based upon the results of the individual t-tests of the coefficients in the combined model. Only one of the coefficients from the structural realist model is significant, and it is in the wrong direction. On the other had, all of the coefficients from the rational deterrence model are in the correct direction, and six of eight are conventionally significant. Based on Table 1, it appears that Huth and his co-authors' conclusions are warranted.

Regime Type and Initiation versus Realism

Do regime type and war initiation offer a better explanation of war outcomes than realist variables? This question is not central to Reiter and Stam's (1998b) article on democracy and victory. Reiter and Stam's actual interest is in tracing the effect of political structure on war outcomes; that is, they are interested in determining why democracies win more wars than nondemocracies. Is it because democracies are intrinsically more effective at waging war, or because democracies are more careful about the decision to initiate conflict? The argument in favor of the former is that it is easier for democracies to rally their society behind a war effort and that democratic armies fight "with greater initiative and better leadership than do the armies of other kinds of states"

[1]Signorino and Yilmaz (2000) demonstrate mathematically why traditional models are misspecified in the strategic case.

[2]Huth, Gelpi, and Bennett (1993) actually test five separate models of structural realism. I have focused solely on their most comprehensive model.

Table 1 A Probit Model of Great Power Dispute Escalation

Variable	Coefficient	S.E.[a]	Significance
Constant	–0.71	(1.32)	
Structural Realism			
System uncertainty 1 (size)	0.21	(0.34)	
System size*risk	–0.97	(0.31)	$p < 0.005$
System uncertainty 2 (diffusion)	–0.20	(0.22)	
System diffusion*risk	0.18	(0.29)	
Risk-acceptant	1.55	(1.32)	
Deterrence Theory			
Balance of forces	1.73	(0.84)	$p < 0.05$
Secure 2nd strike	–2.33	(0.73)	$p < 0.005$
Defender vital interests	–1.29	(0.40)	$p < 0.005$
Challenger vital interests	1.09	(0.39)	$p < 0.01$
Defender backed down	1.23	(0.49)	$p < 0.025$
Challenger backed down	–0.72	(0.61)	
Defender other dispute	0.96	(0.32)	$p < 0.005$
Challenger other dispute	0.05	(0.36)	

[a]Reported standard errors are robust standard errors.

(Reiter and Stam 1998b). The argument in favor of the latter is that democratic leaders face greater post-defeat political consequences than do other kinds of states and therefore initiate wars only when the likelihood of victory is high.

In answering this question, Reiter and Stam estimate five models, one which corresponds to a realist model of war outcomes and one which corresponds to a model that reflects the effects of regime type and the decision to initiate war. These models provide an excellent opportunity to test the question that Reiter and Stam do not address directly—whether the realist explanation is better than the nonrealist explanation. The realist argument is that war outcomes are best explained by the distribution of capabilities across combatants, the quality of the militaries involved, strategy choice, terrain, and the effects of allies. The nonrealist argument focuses on the regime type of the combatants and the decision to initiate war. A replication of Reiter and Stam's results is in Table 2.[3]

The Nonnested Nature of the Models

Nonnested models are found throughout the literature of international relations. Recent international relations articles that contain nonnested models include Lai and Reiter (2000), Davenport (1999), Feng and Zak (1999), Palmer and David (1999), Rasler and Thompson (1999), Shin and Ward (1999), Signorino (1999), Smith (1999), Enterline (1998), Morrow, Siverson, and Taberes (1998), Reiter and Stam (1998b), Reiter and Stam (1998a), Bennett (1997), Gelpi (1997), Bennett and Stam (1996), Benoit, Hermann and Kegley (1996), Lemke and Reed (1996), Pollins (1996), Smith (1996), Huth, Gelpi, and Bennett (1993), Huth and Russett (1993), and Maoz and Russett (1993).[4]

In this section, I define the concept of nonnested and provide a loose typology of the nonnested models found in the above articles.

Defining "Nonnested"

Defining the concept of "nonnested" precisely is not an easy task. Definitions are often imprecise and uncomplicated or precise and complicated. To avoid any potential confusion, I first define nonnested in terms that are imprecise but that are easily understood by nonmethodologists. I then define nonnested in precise mathematical terms. Finally, I make the connection between the definitions and provide a final definition that rests somewhere between these two extremes.

[3]I used a logit link function as opposed to the originally reported probit link function in order to be consistent with a later analysis. No inferences are affected.

[4]Recognizing nonnested models in the literature is not always straightforward. Maoz and Russett (1993), for instance, compress a number of alternative explanations for the democratic peace into one or two variables in an attempt to avoid the problem of nonnested models.

Table 2 Logit Models of War Outcomes

	Model 1		Model 2	
Variable	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
Constant			–6.99	(2.486)
Nonrealist				
Politics*Initiation	0.072**	(0.034)		
Politics*Target	0.060**	(0.028)		
Initiation	0.826****	(0.252)		
Realist				
Capabilities			5.671****	(0.903)
Alliance Contributions			6.462****	(1.05)
Quality Ratio			0.160***	(0.09)
Terrain			–15.270****	(4.284)
Strategy*Terrain			4.816****	(1.382)
Strategy			8.817**	(4.286)
Strategy 2			4.502	(3.012)
Strategy 3			4.278*	(2.193)
Strategy 4			3.707*	(2.008)
Log-Likelihood	–128.1		–74.9	

[a]*p < 0.1, **p < 0.05, ***p < 0.01, ****p < 0.001

[b]Reported standard errors are robust standard errors.

In imprecise but easily understood language, we define two models as nested or nonnested based on whether or not one model is a "special case" of the second model.

Definition 1 (**Nested**) *Two models are nested if one model can be reduced to the other model by imposing a set of linear restrictions on the parameter vector.*

Let us take, for example, two models, H_f and H_g, that are characterized by the same functional form and the same error structure. Express the data as deviations from their means so that no intercept appears.[5] These models differ, then, only in terms of their regressors. In the following specification:

$$H_f\colon Y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon_0 \quad (1)$$

$$H_g\colon Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_1 \quad (2)$$

the models are nested because by imposing the restriction that $\beta_3 = 0$, H_g becomes H_f. In other words, H_g "encompasses" H_f. Discriminating between these models involves simply testing the restriction on β_3. This test can be done with a t-test under ordinary least-squares (Greene 1997) or a likelihood-ratio test under maximum likelihood (King 1989). If H_g included a $\beta_4 x_4$ as well:

$$H_f\colon Y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon_0 \quad (3)$$

$$H_g\colon Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_1 \quad (4)$$

an F-test or likelihood-ratio test would be appropriate (Greene 1997).

Definition 2 (**Nonnested**) *Two models are nonnested, either partially or strictly, if one model cannot be reduced to the other model by imposing a set of linear restrictions on the parameter vector.*

For example:

$$H_f\colon Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_0 \quad (5)$$

$$H_g\colon Y = \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon_1 \quad (6)$$

are nonnested models because even if we impose the restrictions that $\beta_4 = 0$ and $\beta_5 = 0$, H_g does not become H_f. The above models are *partially* nonnested because they have one variable in common, X_3. If H_f and H_g do not share X_3:

$$H_f\colon Y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon_0 \quad (7)$$

$$H_g\colon Y = \beta_4 x_4 + \beta_5 x_5 + \varepsilon_1 \quad (8)$$

the models are strictly nonnested.

Technical definitions of nonnested center around a statistical measure of the "closeness" between two models

[5] This move is made only for pedagogical reasons.

called the Kullback-Leibler information criteria (KLIC). When comparing two models, H_f and H_g, the KLIC is defined as:

$$\int_R \ln\left\{\frac{f(y,\theta)}{f(y,\gamma)}\right\} f(y,\theta)dy$$

where R is the range of variation of Y under H_f. The KLIC is the mean information for discrimination in favor of $f(y, \theta)$ against $g(y, \gamma)$ (Kullback 1959). The measure is interpreted as the surprise experienced on average when we believe that $f(y, \theta)$ is the data generating process (DGP) and then we find that $g(y, \gamma)$ is the DGP (White 1994). The KLIC is used because of its analytic tractability and important properties: the KLIC is invariant to transformations of θ and γ, is nonnegative, is additive for independent random events, and equals 0 when $f(y, \theta)$ and $g(y, \gamma)$ coincide (Kullback 1959; Pesaran 1987).

Using the KLIC allows us to define two models as nested, partially nonnested or strictly nonnested. Following Pesaran (1987), let θ_0 be the true value of θ under H_f and let γ_0 be the true value of γ under H_g. The Kullback-Leibler information criteria for the discrimination of $f(y, \theta)$ against $g(y, \gamma)$ is:

$$I_{fg}(\theta, \gamma) = E_0\{\ln f(y, \theta) - \ln g(y,\gamma)\}$$

which has a unique minimum at $\gamma_*(\theta_0)$. This last quantity is a "pseudo-true" value which means that it is the value that γ would take were $f(y, \theta)$ the true DGP. The closeness of H_g to H_f is then:

$$C_{fg}(\theta_0) = I_{fg}\{\theta_0, \gamma_*(\theta_0)\}$$

We are now in a position to define nested, partially nonnested, and strictly nonnested in terms of the closeness of two models.

Definition 3 (Nested) *Model H_f is nested within model H_g if and only if $C_{fg}(\theta_0) = 0$ for all admissible values of θ_0. Similarly, model H_g is nested within model H_f if and only if $C_{gf}(\gamma_0) = 0$ for all admissible values of γ_0.*

Definition 4 (Strictly Nonnested) *Models H_f and H_g are strictly nonnested if $C_{fg}(\theta_0)$ and $C_{gf}(\gamma_0)$ are both nonzero for all admissible values of θ_0 and γ_0.*

Definition 5 (Partially Nonnested) *Models H_f and H_g are partially nonnested if $C_{fg}(\theta_0)$ and $C_{gf}(\gamma_0)$ are both nonzero for some but not all admissible values of θ_0 and γ_0.*

The connections between Equations (1–8) and their relevant categories defined in terms of the KLIC and closeness are easy to see. Let us assume that Equation (1) is the true model. The pseudo-true value of β_3 in Equation (2), assuming that Equation (1) is true, is zero because β_3 does not appear in Equation (1). The other terms in the models are the same. The closeness of these models is then:

$$C_{fg}(\theta_0) = I_{fg}\{\theta_0, \gamma_*(\theta_0)\} = E_\theta\{\ln f(y, \theta) - \ln g(y, \gamma_*(\theta_0))\} = 0.$$

The models are therefore nested.

In contrast, assume that Equation (5) is the true model. Again, the pseudo-true values of β_4 and β_5 in Equation (6) are zero because neither appear in Equation (5). The other terms in the models, however, are not the same. The closeness of these models is then:

$$C_{fg}(\theta_0) = I_{fg}\{\theta_0, \gamma_*(\theta_0)\} = E_\theta\{\ln f(y, \theta) - \ln g(y, \gamma_*(\theta_0))\} \neq 0.$$

Now let us assume that Equation (6) is true. The pseudo-true values of β_1 and β_2 in Equation (5) are zero because neither appear in Equation (6). Again, the other terms in the model are not the same. The closeness of these models is then:

$$C_{fg}(\gamma_0) = I_{gf}\{\gamma_0, \theta_*(\gamma_0)\} = E_\gamma\{\ln g(y, \gamma) - \ln f(y, \theta_*(\gamma_0))\} \neq 0.$$

As $C_{fg}(\theta_0) \neq 0$ and $C_{gf}(\gamma_0) \neq 0$ for all values of θ_0 and γ_0, the models are strictly nonnested.

Care must be taken when discussing the strictly nonnested case. There are two cases where models with different sets of regressors may not be strictly nonnested. The first is the trivial case where either θ_0 or γ_0 equals zero, in which case $C_{fg}(\theta_0) = 0$ and $C_{gf}(\gamma_0) = 0$ and the models are then nested. We can ignore this case in most instances. The second case is where one or more explanatory variables in one model may be written as a linear combination of the explanatory variables in the second model, in which case $C_{fg}(\theta_0) = 0$ or $C_{gf}(\gamma_0) = 0$ for some or all values of θ_0 or γ_0 (Pesaran 1987). If no variables in the first model can be written as a linear combination of the variables in the second model, the models are strictly nonnested. If one or more, but not all, of the explanatory variables in the first model can be written as a linear combination of the explanatory variables in the second model, the models are partially nonnested. If all the explanatory variables in the first model can be written as a linear combination of the variables in the second model, the models are nested.

As implied in Equations (5–6) and (7–8), whether or not the models are partially or strictly nonnested does not matter; the problems, and the solutions to those problems, are the same. We are now in a position to provide a compromise definition.

Figure 1 A Typology of Nonnested Models in International Relations

	Comparative Testing	Robustness Checking
Functional form	Signorino (1999) Smith (1999)	Lai and Reiter (2000) Shin and Ward (1999) Smith (1999) Benoit (1996) Smith (1996)
Covariates	Palmer and David (1999) Reiter and Stam (1998b) Reiter and Stam (1998a) Bennett (1997) Gelpi (1997) Bennett and Stam (1996) Pollins (1996) Huth, Gelpi, and Bennett (1993) Huth and Russett (1993) Maoz and Russett (1993)	Lai and Reiter (2000) Davenport (1999) Feng and Zak (1999) Rasler and Thompson (1999) Enterline (1998) Morrow, Siverson, and Taberes (1998) Benoit (1996) Hermann and Kegley (1996) Lemke and Reed (1996) Smith (1996)

Definition 6 (Final) *Two models with the same functional form and the same error structure are nonnested* if and only *if at least one explanatory variable in each regressor matrix cannot be written as a linear combination of the explanatory variables of the other model.*

From Table 1, it should be clear that the rival models being tested by Huth, Gelpi, and Bennett (1993) are nonnested and are, in fact, strictly nonnested. An easy test of whether or not the models are strictly nonnested is to combine both regressor matrices into a single matrix and attempt to invert it. If the combined regressor matrix inverts, it must be of full rank, which means that each column is linearly independent of the others. The two models in Table 1 have no variables in common, the combined regressor matrix inverts, and therefore the closeness of the models does not equal zero.[6] The same characteristics hold true for the Reiter and Stam models in Table 2: the models have no variables in common, the combined regressor matrix inverts, and therefore the models are strictly nonnested.

A Typology

Articles in political science that estimate nonnested models may be classified along two dimensions: the reason why the authors estimated nonnested models and the reason why the models are nonnested. Along the first dimension, nonnested models are used either to test rival theories or as robustness checks on a single theory. Along the second dimension, models are nonnested either in terms of their functional form or their covariates. Nonnested analyses fall, therefore, into one (or more) of four categories: comparative/functional, comparative/covariates, robustness/functional, and robustness/covariates. Figure 1 lists the various combinations as well as recent international relations articles that fall into each cell. For example, Smith (1999) is in the comparative/functional cell as he compares a bivariate-ordered probit model to a strategically censored discrete choice model. These models are not simply different specifications, but different theories. Shin and Ward (1999), on the other hand, estimate the same theory of military spending on economic growth using both a spatial lag specification and a spatial error specification. This article falls into the robustness/functional cell. Pollins (1996) estimates nonnested models in an attempt to compare long cycle theories of international relations. These models are nonnested only in terms of their covariates so the article falls into the comparative/covariates cell. Finally, Hermann and Kegley (1996) use nonnested models to trace assess the effect of polity type on being a target of intervention across a number of specifications, all of which have the same functional form. The article is then classified under robustness/covariates.

The categories in Figure 1 are not mutually exclusive. An article in the comparative/functional cell is likely to fall into the comparative/covariates cell as well. Some, such as Lai and Reiter (2000), check the robustness of their results across covariates and functional forms and hence fall into two categories. In a field where many articles report estimates from a large number of models, a mix of nested and nonnested models in the same article is quite common.[7]

[6] We can safely reject the possibility that all the parameters in either model equal zero.

[7] The average number of models reported in the articles in Figure 1 is six.

The purpose of presenting this typology is to justify the focus of this article on models that are nonnested in terms of their covariates. As Figure 1 attests, comparative theory testing or even robustness checking of models with different functional forms has yet to enter the methodology of international relations in a significant manner. Competing models in the published literature generally differ only in terms of their covariates. Of the twenty-three articles in Figure 1, seventeen concern models that are nonnested only in terms of their covariates.

I also focus my comments on the generalized linear model, particularly those models where the link function is nonlinear. The reason for this choice is that statistical models in international relations generally have discrete dependent variables. I specifically look at the case where the link function is a probit or logit, but the methods I discuss are easily extended to other link functions. Discriminating between nonnested linear models is easily accomplished with Davidson and MacKinnon's "J" and "JA" tests (see Davidson and MacKinnon 1993; Judge et al. 1985; Kmenta 986; Greene 1997).

Some Problematic Approaches

Tables 1 and 2 reveal a marked difference in the approach Huth, Gelpi, and Bennett (1993) take to dealing with the problem of testing nonnested rival models and the approach Reiter and Stam take. In Table 1, Huth and his coworkers employ the "supermodel" approach, which requires artificially nesting the two models in a single equation and treating the rival model as a control. In Table 2, Reiter and Stam choose to estimate the models separately. Neither of these approaches, however, is sufficient to compare nonnested models accurately.

Techniques for Artificially Nested "Supermodels"

The supermodel approach consists of combining the variables in the rival models into a single large equation. Inference is based on individual or joint significance tests of the artificially nested models. Artificially nested models can be found in Bennett (1997), Gelpi (1997), Bennett and Stam (1996), Huth and Russett (1993), and Maoz and Russett (1993). For the Huth, Gelpi, and Bennett analysis in Table 1, six of eight coefficients from the rational deterrence model are conventionally significant.[8] In contrast, only one of five coefficients from the structural realist model is significant. The conclusion that Huth, Gelpi, and Bennett draw from these results is that the rational deterrence model receives more support from the data than the structural realist model.

The persuasiveness of this conclusion rests on two questionable techniques. The first is t-tests of individual coefficients within the artificially nested model. The second is controlling for alternative theories within the same equation. These techniques for artificially nested models, along with one not used by Huth, Gelpi, and Bennett, the F-test, are considered below.[9]

T-tests. An obvious problem with using t-tests of individual coefficients to test a model is that single variables rarely encapsulate entire theories. A researcher must then contend with a number of different t-tests within each model, not all of which are likely to agree. No metric exists for assessing how many coefficients should be insignificant before rejecting a theory. In addition, individual t-tests do not generate probabilistic statements regarding model choice. The results in Table 1 provide no sense of the uncertainty surrounding the choice of rational deterrence as the superior model.

Individual significance tests, furthermore, do not always give the same answer as a joint significance test (such as an F-test) would. It is quite possible for a number of variables to be individually insignificant and yet jointly significant (Greene 1997).[10] The fact that only one of five coefficients from the structural realist model is significant cannot be interpreted to mean that the structural realist model has no bearing on the escalation of great-power militarized disputes. In the nonnested case, unfortunately, a simple application of the F-test or a likelihood-ratio test will not solve the problem (see below and the next section).

Finally, when rival models are combined in a single equation, there is often a temptation to compare coefficients in order to identify the "important" ones. The model with the greatest number of important variables is then selected. Comparative claims, however, cannot be based on the relative size of the coefficients. In general, the size of two coefficients cannot be compared (King 1986). The effect on the dependent variable from a unit, percent, or standard deviation change in "system uncertainty 2 (diffusion)" cannot be compared with the effect of a unit, percent, or standard deviation change in "secure second strike." The variables are simply measured on different scales. The best we can do is state whether or

[8]P-value ≤ 0.05

[9]Collinearity is also a potential problem with artificially nested models (see Greene 1997).

[10]In rare cases, it is also possible for a set of coefficients to be individually significant and jointly insignificant.

not each individual variable has an effect on the dependent variable (Achen 1982).

Controlling for alternative theories. The supermodel approach to testing nonnested models arises naturally from the belief that we must control for alternative theories within the same regression model. The argument is that we cannot get a good estimate of the effect of theory one (T_1) if we cannot control for the effects of theory two (T_2). This argument, however, is flawed. Theory dictates the choice of covariates in a regression. Statistical analysis divorced from theory, in Achen's (1982, 15) terms, "is likely to degenerate into ad hoc curve fitting. . . ." Unless there exists a well-specified theory that claims that both T_1 and T_2 are relevant to the phenomena in question, T_1 and T_2 do not belong in the same equation.

Supermodels are, in effect, atheoretical. Including both theories in the same equation is a misspecification equivalent to including irrelevant variables, which affects the precision of the estimates. As of this date, there exists no theory of the escalation of great-power militarized disputes that claims a role for variables from both structural realism and rational deterrence theory. If a combined theory did exist, it would be interesting to test the structural realist model and the rational deterrence model against the combined model.[11] Until such a theory exists, however, these two sets of covariates should remain separate.

Accepting the premise that theory must determine the covariates of a regression does not mean that there is no place for nonnested hypothesis testing. Which theory out of a group of plausible theories does the best job of explaining a phenomenon remains a question. This is the question that nonnested tests help us answer.

F-tests. Whether or not the F-test can be used to discriminate between nonnested models depends upon whether or not the models are partially or strictly nonnested. The F-test cannot be used with partially nonnested models. Consider the following the partially nonnested models in matrix form:

$$H_f: Y = X\beta + \varepsilon_0 \quad (9)$$

$$H_g: Y = Z\gamma + \varepsilon_1 \quad (10)$$

If we artificially nest these models into a single equation where $\tilde{X}$ are the variables in X but not in Z, $\tilde{Z}$ are the variables in Z but not in X, and W are the variables in two models have in common,

$$H_C: Y = \tilde{X}\beta + \tilde{Z}\gamma + W\sigma + \varepsilon$$

we see that the common variables (W) create a problem. Testing either β or γ does not test the full models. On the other hand, testing β and σ or γ and σ 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).

The F-test can be used with strictly nonnested models such as the Huth, Gelpi, and Bennett models in Table 1. If Equations (9) and (10) are nonnested, the combined model is then:

$$H_C: Y = \tilde{X}\beta + \tilde{Z}\gamma + \varepsilon$$

As the models have no common variables, the problem noted in the previous paragraph does not arise. The F-test is rarely used, however, because monte carlo simulations have demonstrated that it has low power (the probability of rejecting a false null) in this situation (McAleer 1987).

Techniques for Separate Models

An alternative approach to testing nonnested models, one taken by Reiter and Stam, is to estimate each model separately and then compare the log-likelihoods. Similar approaches compare the F-tests or likelihood-ratio tests that are reported by most standard software packages. Like the supermodel approach, probabilistic statements regarding model discrimination cannot be made on the basis of these techniques.

Likelihoods. Reiter and Stam estimate their models separately, thus generating two log-likelihood statistics. The log-likelihood of the realist model is larger than the log-likelihood of the nonrealist model. Exactly what to make of this fact is unclear. The number of coefficients in the models affect the log-likelihoods. The greater the number of coefficients, the greater the log-likelihood. In a nested model, we could use a likelihood-ratio test where the difference in the number of coefficients would be handled by the degrees of freedom of the χ^2-distribution.[12] Reiter and Stam's models, however, are nonnested. The realist model has seven more coefficients than the nonrealist model and part of the difference in the log-likelihoods must be due to the extra included variables. We cannot be sure how much. Furthermore, the fact that we cannot

[11] Such a test would be a nested test.

[12] Likelihood-ratio tests may only be used with nested models. This point is not universally appreciated. Pollins (1996) incorrectly uses likelihood-ratio tests to discriminate between nonnested long-cycle theories in an otherwise well-regarded article.

perform a likelihood-ratio test means that no probabilistic statement can be made on the basis of the log-likelihoods. Without such a statement, we cannot say how certain we are that one likelihood is significantly larger than the other.

Likelihood-ratio tests and F-tests. Though not common, model discrimination statements are occasionally made on the basis of likelihood-ratio tests and F-tests where the null is a completely restricted model. These tests are commonly reported by standard software packages. The smaller the p-value of these tests, the argument goes, the better the model fits the data. There is nothing about these tests and their associated p-values, however, that supports such a conclusion. The idea of Neyman-Pearson hypothesis testing is to use the information contained in the p-value in a decision-making context. If the p-value is less than the *a priori* probability of a type I error, the null hypothesis is rejected. No conclusions regarding the alternative hypothesis are to be drawn. A point that is not well understood about this procedure is that how much smaller the p-value is relative to the significance level is immaterial. P-values do not provide measures of support for or against hypotheses. In classical (as opposed to Bayesian) statistics, hypotheses are not probabilistic. The null hypothesis is true with probability 0 or 1. There is no theory, in the Neyman-Pearson system, that connects the significance level of a hypothesis test with a measure of inductive support for that hypothesis (Howson and Urbach 1993).[13]

Rejecting the null hypothesis in an F-test or likelihood-ratio test implies that chance or randomness is not a good explanation for the phenomenon being explained. This is a useful result if the goal is to explain a phenomenon where chance is a powerful theory. If we are able to reject chance as a model for the numbers that come up in roulette, where chance is a powerful explanation, we have learned something important. Chance is not a powerful explanation, however, for the type of phenomena that international relations scholars attempt to explain. We therefore do not learn much by rejecting the null hypothesis with these tests.[14] The great strength of the nonnested hypothesis tests presented in the next section is that if we reject the null hypothesis, we are rejecting a hypothesis of substantive interest.

[13] This does not imply that regression is useless. The main point of regression is the estimation of β.

[14] If, on the other hand, we do not reject the null, we have learned that the model is no better than chance.

Tests for Nonnested Models

In this section, I present three appropriate tests for discriminating between nonnested rival models.[15]

The Cox Text

The literature on nonnested model testing stems from the seminal work of David Cox (1961, 1962). The philosophical underpinning of Cox's test is that a true model should be able to predict the performance of specific alternatives. The idea is to compare the actual performance of the alternative model with the expected performance of the alternative model under the null hypothesis (McAleer 1987). A true null should not distort the actual performance of the alternative model.

Econometricians have primarily used the Cox test to distinguish between models with different functional forms, and the test is particularly powerful in this situation. On the other hand, any social scientist looking up "selection of regressors" in a major econometrics text will find the Cox test. Greene (1997), Kennedy (1992), Davidson and MacKinnon (1993), Kmenta (1986), and Judge et al. (1985) all associate the Cox test, particularly in its linear form, with models that are nonnested in terms of their covariates. Of the three methods discussed in this article, the Cox test is the only one that appears in every one of the texts listed above.

The math behind Cox's innovation is a generalization of the familiar likelihood-ratio test statistic. The modified statistic is the difference between the log-likelihood ratio and the expected log-likelihood ratio under the null hypothesis.[16] That is, if $L_f(\hat{\theta}_f)$ is the maximum value of the likelihood of a sample of y values when H_f is postulated and $L_g(\hat{\gamma}_g)$ is the maximum value of the likelihood of a sample of y values when H_g is postulated, then the log of the likelihood ratio is:

$$\hat{l}_{fg} = \ln L_f(\hat{\theta}_f) - \ln L_g(\hat{\gamma}_g) \tag{11}$$

The numerator of the Cox statistic is the difference between Equation (11) and the expected log-likelihood ratio under the null hypothesis:

$$T_f = \hat{l}_{fg} - E(\hat{l}_{fg}) \tag{12}$$

[15] Other discrimination procedures for nonnested models exist in the literature (see Horowitz 1983; Davidson and MacKinnon 1993). The three methods detailed here are the most common.

[16] Both models must, in turn, serve as the null hypothesis.

The Cox statistic is then the difference in Equation (12) suitably normalized:

$$N_f \equiv \frac{T_f}{\left[V\left(T_f\right)\right]^{1/2}} \sim N(0,1).$$

Cox's statistic is sometimes referred to as a centered-likelihood ratio, as N_f has a standard normal distribution.

The major problem with applying the Cox statistic to models other than least-squares models is calculating the expected value of the log-likelihood ratio under the null hypothesis. We can approximate this expected value using the Kullback-Leibler information criteria, the measure of closeness defined in an earlier section. Where the KLIC cannot be analytically derived, a simulation approach is necessary (Pesaran and Pesaran 1993). For the binary choice models we are using, an analytical derivation of the KLIC is possible, and we need only simulate what statisticians refer to as the pseudo-maximum likelihood estimator—a consistent estimator produced by a misspecified model (see White 1994; Gourieroux and Monfort 1995) .

Following the approach used by Pesaran (1987) and Pesaran and Pesaran (1993), the numerator of the test statistic is:

$$L_f(Y,\hat{\theta}) - L_g(Y,\hat{\gamma}) - C(\hat{\theta},\hat{\gamma}_\star(R))$$

where $C(\hat{\theta},\hat{\gamma}_\star(R))$ is the estimated KLIC. Notice that $\hat{\gamma}_\star$ is the maximum-likelihood estimator of H_g assuming that H_f is the actual data generating process. $\hat{\gamma}_\star$ is therefore a pseudomaximum-likelihood estimator and R stands for the number of repetitions used to simulate the estimator. Running the simulations is not difficult; a three-step procedure is outlined in Figure 2.

FIGURE 2 Running the Simulations

Step One: Run the null model, H_f, against the observed *y*-vector and save the estimated coefficient vector, $\hat{\theta}_f$.

Step Two: Produce a simulated *y*-vector, $\hat{y}$, using $\hat{\theta}_f$ and the data.

Step Three: Run H_g, the alternative model, on the simulated *y*-vector, $\hat{y}$. The resulting estimates, $\hat{\gamma}_\star$ are pseudo-maximum likelihood estimates because they are produced by an assumed misspecified model run on a dependent variable generated by the null.

Step Four: Replicate this estimation and average over the results.

The standard error of the statistic is :

$$N^{-1}d'\left\{I_N - R(\hat{\theta})[R'(\hat{\theta})R(\hat{\theta})]^{-1}R'(\hat{\theta})\right\}d$$

where d is the observed-likelihood ratio and $R(\hat{\theta})$ is a matrix of partial derivatives of $\ln(Y,\theta)$ with respect to θ. See Appendix A for the math worked out in terms of rival binary-choice models.

In very simple terms, the Cox test resembles a χ^2 test in form. The χ^2 test is based on the difference between an observed value and an expected value. The general idea of the Cox test is similar:

$$\frac{\text{Observed log-likelihood ratio} - \text{Expected log-likelihood ratio}}{\text{S.E.}} = 0$$

if the null hypothesis is true. We reject the null hypothesis if the test statistic differs significantly from zero in either direction.

The results of a Cox test are not always unambiguous. As there is generally no reason to assume that either rival model is the null, each model must, in turn, take that role. Four outcomes are therefore possible: one or the other model may be rejected, both models may be rejected, or neither model may be rejected. Just as in any hypothesis test, it is important to remember that if the null is rejected, it is not rejected in favor of the alternative.

The possibility of rejecting both models without a hint of what to do next has led some to criticize the Cox test (Granger, King, and White 1995). The fact that both models may be rejected or neither model rejected should not, however, be taken as a weakness of the test; the rejection of both models implies that neither model could predict the results of the other model. We should conclude, then, that both models are misspecified in some way. Such a result is not inconsistent with the results of individual likelihood-ratio tests against a completely restricted model. Even misspecified models may be strong enough to reject a null hypothesis of no effect.

The Vuong Test

What to do if both models are rejected is, of course, a natural question. It is in this situation that Vuong's (1989) model selection test proves useful. Short of inventing better theories, we might want to choose the best of a bad lot of models and work at respecifying that model. Similarly, if we fail to reject either model, we might use model selection criteria such as Vuong's to identify the model that is closer to the true specification.

Vuong's test also makes use of the Kullback-Leibler information criteria. Vuong defines the KLIC as:

$$KLIC \equiv E^0[\ln h^0(Y_t|X_t)] - E^0[\ln f(Y_t|X_t;\theta_*)] \quad (13)$$

where $h^0(.|.)$ is the true conditional density of Y_t given X_t (that is, the true but unknown model), and θ_* are the pseudotrue values of θ (the estimates of θ when $f(Y_t|X_t)$ is not the true model). The best model is the model that minimizes Equation (13), for the best model is the one that is closest to the true specification. We should therefore choose the model that maximizes $E^0[\ln f(Y_t|X_{tj}\theta_*)]$. 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:

$$H_0: E^0\left[\ln\frac{f(Y_t|X_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] = 0$$

meaning that the two models are equivalent.[17] The alternative hypotheses are:

$$H_f: E^0\left[\ln\frac{f(Y_t|X_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] > 0$$

$$H_g: E^0\left[\ln\frac{f(Y_t|X_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] < 0$$

meaning that H_f is better than H_g or H_f is worse than H_g, respectively.

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

$$\frac{1}{n}LR_n(\hat{\theta}_n,\hat{\gamma}_n) \xrightarrow{a.s.} E^0\left[\ln\frac{f(Y_t|X_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]$$

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

$$\text{under } H_0: \frac{LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} N(0,1) \quad (14)$$

$$\text{under } H_f: \frac{LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} +\infty \quad (15)$$

$$\text{under } H_g: \frac{LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} -\infty \quad (16)$$

where

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

and,

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

Like the Cox test, 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-likelihood ratio used in Equations (14–16) can be affected if the number of coefficients in the two models being estimated is different. Unlike the Cox test, the Vuong test needs a correction for the degrees of freedom.[18] The adjusted statistic is:

$$\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n) \equiv LR_n(\hat{\theta}_n,\hat{\gamma}_n) - K_n(F_\theta, G_\gamma)$$

where $K_n(F_\theta, G\gamma)$ is the correction factor. Vuong (1989) suggests using a correction that corresponds to either Akaike's (1973) information criteria or Schwarz's (1978) Bayesian information criteria.[19] I have chosen the latter, making the adjusted statistic [20]:

[17] γ_* and Z in model g are analogous to θ_* and X in model f.

[18] The Cox test does not need a correction as it uses the log-likelihood ratio and the *expected* log-likelihood ratio, which has already taken the degrees of freedom into account.

[19] 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

FIGURE 3 Calculating the Vuong Test for the Huth, Gelpi, and Bennett (1993) Models

Step One: Run H_f and save the reported log-likelihood (–55.933).

Step Two: Run H_g and save the reported log-likelihood (–45.558).

Step Three: Calculate the degrees of freedom correction for the two models:

$$\left|\left(\frac{5}{2}\right)\ln(97)-\left(\frac{8}{2}\right)\ln(97)\right| = -6.86.$$

Step Four: The numerator is then:

[–55.933 – (–45.558)] – 6.86 = –3.5.

Step Five: The variance is the expectation of the squared difference in the individual log-likelihoods minus the square of the expectation:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{f\left(Y_t\middle|X_t;\hat{\theta}_n\right)}{g\left(Y_t\middle|Z_t;\hat{\gamma}_n\right)}\right]^2-\left[\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f\left(Y_t\middle|X_t;\hat{\theta}_n\right)}{g\left(Y_t\middle|Z_t;\hat{\gamma}_n\right)}\right]^2} = 4.5$$

Step Six: The Vuong statistic then is the value in step four divided by the square root of the value calculated in step five:

–3.5/4.5= –0.78.

$$L\tilde{R}_n(\hat{\theta}_n,\hat{\gamma}_n) \equiv LR_n(\hat{\theta}_n,\hat{\gamma}_n)-\left|\left(\frac{p}{2}\right)\ln n-\left(\frac{q}{2}\right)\ln n\right|$$

where p and q are the number of estimated coefficients in models f and g, respectively. Figure 3 walks through the calculation of the Vuong statistic for the Huth, Gelpi, and Bennett (1993) models and Appendix B works out the math for rival binary choice models.

The difference between the Cox test and the Vuong test lies in the null hypothesis. The Cox test takes one of the models under consideration as the null. The null hypothesis for the Vuong test is that there is no difference between the models. The implication of this difference is that application of the Cox test may result in the rejection of both models. 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 the specification. The distinction is between an absolute test where the models are evaluated against the data (the alternative model provides the power) and a relative test where the models are evaluated against the data and each other.

Bayes Factors

The data set in Huth, Gelpi, and Bennett (1993) consists of the population of great-power extended and direct immediate deterrence encounters from 1816 to 1984. The data set in Reiter and Stam (1998b) is the population of interstate wars between 1816 and 1982. In neither case can we easily argue that we are dealing with a sample. Classical tests procedures such as the Cox and Vuong tests assume that repeated sampling is possible. Unless one resorts to the philosophically weak argument that the world *could* have produced other outcomes, classical tests are inappropriate for these data. The solution is to turn to Bayesian inference.

The Bayesian approach to the problem of discriminating between nonnested models is quite different from the classical approach. Bayesians reject the notion that one model should be accepted to the exclusion of a second model, for neither model is likely to be the truth. Rather, Bayesians use sample data to update their assessment of the comparative likelihood of the two models (Greene 1997). A nice international relations application of Bayes factors for nonnested models can be found in Smith (1999).

Following Raftery (1995), let D be the data and M_f and M_g be two nonnested models we wish to compare with parameter vectors θ_f and θ_g. If $P(M_f)$ is the prior probability of model f and $P(D|M_f)$ is the integrated likelihood of model f, the posterior probability of M_f is given by Bayes theorem:

$$P\left(M_f\middle|D\right) = \frac{P\left(D\middle|M_f\right)P\left(M_f\right)}{P\left(D\middle|M_f\right)P\left(M_f\right)+P\left(D\middle|M_g\right)P\left(M_g\right)} \quad (17)$$

where,

$$P\left(D\middle|M_f\right) = \int P\left(D\middle|\theta_f,M_f\right)P\left(\theta_f\middle|M_f\right)d\theta_f$$

The posterior probability of M_g is:

$$P\left(M_g\middle|D\right) = \frac{P\left(D\middle|M_g\right)P\left(M_g\right)}{P\left(D\middle|M_f\right)P\left(M_f\right)+P\left(D\middle|M_g\right)P\left(M_g\right)} \quad (18)$$

et al. 1985). These measures are particularly useful in a time-series framework where forecasting is the goal (Greene 1997). Like the Vuong test, but unlike the Cox test, these measures will always choose a model even if neither model fits the data well (McAleer 1987). Unlike either the Vuong or the Cox test, these techniques do not provide probabilistic statements regarding model selection and neither technique incorporates knowledge from the rival model.

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

Notice that the denominators of Equations (17) and(18) are the same. We can therefore take the ratio of the two posteriors:

$$\frac{P(M_g|D)}{P(M_f|D)} = \left[\frac{P(D|M_g)}{P(D|M_f)}\right]\left[\frac{P(M_g)}{P(M_f)}\right]. \quad (19)$$

The first factor on the right-hand side of (19) is defined as the Bayes factor (the ratio of integrated likelihoods), while the second factor is the prior odds ratio. Equation (19), then, corresponds to:

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}. \quad (20)$$

Assuming that we assign equal priors to the models, the prior-odds ratio equals 1 and drops out. The posterior-odds ratio then equals the Bayes factor. When the Bayes factor is greater than 1, the data favor M_g over M_f. When the Bayes factor is less than 1 the reverse is true. Raftery (1995), following Jeffreys (1961), proposes the following "rules of thumb" for interpreting twice the logarithm of the Bayes factor:

$0 \le 2 \ln(BF) \le 2.2$	Very weak evidence for M_g
$2.2 \le 2 \ln(BF) \le 5$	Weak to moderate evidence for M_g
$5 \le 2 \ln(BF) \le 10$	Moderate to strong evidence for M_g
$2 \ln(BF) > 10$	Decisive evidence for M_g

Posterior probabilities for each model can be calculated from the Bayes factors of each model against a null model. If M_i are the models being compared with the null (M_0), B_{i0} are the corresponding Bayes factors, and α_i are the prior odds, then:

$$P(M_i|D) = \frac{\alpha_k \beta_{k0}}{\sum_{i=0}^{K} \alpha_i \beta_{i0}} \quad (21)$$

are the posterior probabilities of the models given the data.

While the theory and use of Bayes factors are straightforward, calculating the integrated likelihoods necessary for Bayes factors is not straightforward. Three main approaches to calculating Bayes factors exist in the literature. The first is a relatively simple version of the Laplace approximation known as a BIC (Bayesian information criteria) approximation. This is the approach taken by Bartels (1997). The second is a Markov chain Monte Carlo approach used by Smith (1999) and Clarke (2000). The third approach, and the one used here, is a more complicated, but more accurate, version of the Laplace approximation. Jeffreys (1961) first proposed using the Laplace method for approximating Bayes factors, and Raftery (1993) applied the method to generalized linear models.[21]

Assuming the ratio of the priors equals 1, the approximation provides the first factor in Equation (19):

$$\left[\frac{P(D|M_g)}{P(D|M_f)}\right].$$

Remember, though, that $P(D|M_i)$ also includes a prior, $P(\theta_i|M_i)$:

$$P(D|M_i) = \int P(D|\theta_i, M_i) P(\theta_i|M_i) d\theta_i$$

The work in estimating a Bayes factor lies in choosing this prior. For generalized linear models, Raftery (1993) argues for a multivariate normal prior with all the parameters except the intercept centered at zero and all the covariances not involving the intercept set equal to zero. The prior then involves only one hyperparameter to which the results are sensitive. This hyperparameter, denoted by φ, controls the prior standard deviations of the regression parameters. To ensure that the prior does not contribute much evidence in favor either model in the nonnested case, φ should be large. On the other hand, large values of φ tend to favor simpler models. In cases of ignorance, Raftery (1993) recommends evaluating the Bayes factors over the range $1 \le \varphi \le 5$ with $\varphi = 1.65$ as a "central" value.[22]

For a specific set of models, it is possible to translate φ into actual standard deviations (see Clarke 2000). For the Reiter and Stam models, setting φ to 1 is roughly equivalent to setting the prior standard deviations of the coefficients to 10, setting φ to 1.65 is roughly equivalent to a prior standard deviation of 32, and setting φ to 5 is roughly equivalent to a prior standard deviation of 100.[23] I report results for $\varphi = \{1, 1.65, 5\}$.

[21] The details of the Laplace method for integrals is beyond the scope of this paper. See Raftery (1993) for details of the method. The estimates were produced with Raftery's "glib" software, an S-Plus program which uses the Laplace method. "Glib" is freely available from Statlib.

[22] The simple BIC approximation (Raftery 1995) sets the prior covariance matrix at i_j^{-1}, where i_j is the expected Fisher information for one observation (Bartels 1997).

[23] The effect of φ would disappear if the sample size were large enough.

TABLE 3 The Cox Test for the Huth, Gelpi, and Bennett (1993) Models

		Model One as Null[a]		Model Two as Null	
Sim #	n	Z Stat	Significance	Z Stat	Significance
1	300	–0.6747	0.4999	–0.1028	0.9181
2	400	–0.6727	0.5011	–0.1019	0.9189
3	500	0.6757	0.4992	–0.1042	0.9179
4	1000	–0.6739	0.5004	–0.1016	0.9190

[a]Model one is the structural realist model.

TABLE 4 The Vuong Test for the Huth, Gelpi, and Bennett (1993) Models

Vuong	Std. Error	Z Stat	Significance	95% Confidence Int.	
–3.5	4.5	–0.78	0.435	–12.33	5.31

Results

We are now in a position to evaluate the rival nonnested models presented by Huth, Gelpi, and Bennett and Reiter and Stam. The results of the Cox test applied to the structural realist and rational deterrence models are in Table 3. Both null models, according to the Cox test, are accepted at conventional significance levels.[24] The test therefore fails to discriminate between the models. Notice, however, that the test statistic when the structural realist models serves as the null is further from zero than when the rational deterrence model is the null. The results of the test therefore lean slightly in favor of the rational deterrence model. This evidence is quite weak, though. The conclusion, then, is that a result that appeared to be clear-cut in the Huth, Gelpi, and Bennett analysis is now problematized. Our level of uncertainty regarding the explanatory power of the rational deterrence model over the structural realist model has increased dramatically.

The results of the Vuong test are in Table 4. Like the Cox test, the Vuong test fails to discriminate between the models. We cannot place much confidence in the location of the test statistic given the relative size of the standard error. Far from the strong support that Huth, Gelpi, and Bennett find for the rational deterrence model, the results indicate that the evidence in favor of the rational deterrence model is weak and uncertain.

There is no mystery why neither the Cox test nor the Vuong test can discriminate between the structural realist and deterrence theory explanations. The answer lies in the correlation between the models. Canonical correlation is related 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 two sets of variables are used to create linear combinations, which are then weighted so that the correlation between the combinations is as high as possible.[25] The canonical correlation between the structural realist model and the rational deterrence model is 0.88, which implies that the structural realist model and the rational deterrence model are highly correlated. The higher the correlation between the models, the less likely it is that either the Cox test or the Vuong can discriminate between them (see Clarke 1999 for monte carlo simulation results).

The Bayesian analysis echoes the results of the Cox and Vuong tests. Due to the high correlation between the models and the relatively small sample size ($n = 97$), the Bayes factors and posterior model probabilities calculated from Equation (21) turn out to be highly sensitive to the choice of φ. The results are in Table 5. As φ moves from 1 to 5, the results move from weak evidence in favor of the rational deterrence model to strong evidence in favor of the structural realist model. The prior would only have this effect if the models were close to indistinguishable in the first place. If we look solely at the "central"

[24] The p-value for each test is greater than 0.05. We therefore make a decision to not to reject both null hypotheses.

[25] Think of canonical correlation analysis as multivariate regression. Instead of regressing a single-response variable on a *set* of covariates, we regress a *set* of covariates on a second set of covariates. Just as OLS regression maximizes the correlation between the linear combination on the right-hand side of the equation and the dependent variable, canonical correlation maximizes the correlation between two linear combinations. The square of the canonical correlation is equivalent to R^2 from a linear least-squares regression.

Table 5 Bayes Factors and Posterior Probabilities for the Huth, Gelpi, and Bennett (1993) Models

	φ = 1	φ = 1.65	φ = 5
Bayes Factors[a]	–3.97	–0.72	6.09
Posterior Probabilities			
Structural Realist Model	0.12	0.41	0.95
Rational Deterrence Theory	0.88	0.59	0.05

[a]Negative values are evidence in favor of the rational deterrence model.

Table 6 The Cox Test for the Reiter and Stam Models

		Model One as Null[a]		Model Two as Null	
Sim #	n	Z Stat	Significance	Z Stat	Significance
1	300	–2.0599	0.0394	–0.1181	0.9060
2	400	–2.0608	0.0393	–0.1181	0.9060
3	500	–2.0592	0.4992	–0.1181	0.9060
4	1000	–2.0598	0.0394	–0.1181	0.9060

[a]Model one is the nonrealist model.

Table 7 The Vuong Test for the Reiter and Stam Models

Vuong	Std. Error	Z Stat	Significance	95% Confidence Int.	
–36.74	8.93	–4.11	0.0	–54.24	–19.24

value of φ, the results very weakly support the rational deterrence model. The "central" value is particularly useful in this situation, for we know that large values of φ tend to favor simpler models. We could have anticipated that the larger value of φ would support the structural realist model as it is the model with fewer parameters.

The models estimated by Reiter and Stam do not pose the same problems as the Huth, Gelpi, and Bennett models. The canonical correlation between the models is 0.58, which means that these models are not as highly correlated as the previous example. The nonnested tests should do a better job of discriminating between these models.

The results of the Cox test are in Table 6, and they are quite clear. When the nonrealist model serves as the null, it is rejected at the 0.05 level. When the realist model serves as the null, the test fails to reject it at conventional significance levels. The realist model can therefore predict the results of the nonrealist model, but the reverse is not true. We conclude, then, that the realist model is consistent with the data and the nonrealist model is inconsistent with the data.

The Vuong test tells the same story, the results of which are in Table 7. Judging from the small standard error, we can have great confidence in the location of the test statistic. The results demonstrate strong support for the conclusion that the realist model is much closer to the true specification than the nonrealist model.

Unlike the results from the last example, the Bayes factors for this example are robust in the face of alternative priors. The Bayesian results are in Table 8. For each value of φ, there is what Raftery (1993) calls "decisive" support for the realist model. That is, for each value of φ, $2\ln(BF)$ is far greater than 10. We can use Equation (21) to go beyond Raftery's "rules of thumb" and actually calculate the posterior probabilities of both models. The posterior probability of the nonrealist model is approximately zero, while the posterior probability of the realist model is approximately one.

Taking the results together, there is little or no evidence in support of the nonrealist model. All three methods indicate that the realist model is the superior explanation of war outcomes.

TABLE 8 Bayes Factors and Posterior Probabilities for the Reiter and Stam Models

	φ = 1	φ = 1.65	φ = 5
Bayes Factors[a]	78.07	73.69	61.52
Posterior Probabilities			
Nonrealist Model	≈ 0	≈ 0	≈ 0
Realist Model	≈ 1	≈ 1	≈ 1

[a]Positive values are evidence in favor of the realist model.

TABLE 9 Five Model Specifications and Posterior Probabilities

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Nonrealist					
Poly-Pol 1*Initiation		x		x	
Poly-Pol 2*Initiation		x		x	
Politics*Initiation	x		x		
Political*Target	x	x		x	x
Initiation	x	x		x	x
Realist					
Capabilities			x	x	x
Alliance Contributions			x	x	x
Quality Ratio			x	x	x
Terrain			x	x	x
Strategy*Terrain			x	x	x
Strategy 1			x	x	x
Strategy 2			x	x	x
Strategy 3			x	x	x
Strategy 4			x	x	x
Posterior Probabilities					
φ = 1	≈ 0	≈ 0	0.11	0.67	0.21
φ = 1.65	≈ 0	≈ 0	0.28	0.56	0.16
φ = 5	≈ 0	≈ 0	0.90	0.09	0.01

Whether the realist or the nonrealist model provides a better explanation of war outcomes was not the question of interest for Reiter and Stam. Their question was whether regime type and initiation have independent effects on war outcomes while controlling for the realist variables. Light can be shed on this question by using Equation (21) and our Bayes factors to produce posterior model probabilities.

The results of Reiter and Stam's tests indicate that regime type and initiation do have independent effects. The specifications for the five models Reiter and Stam estimated are in Table 9.[26] From the first two columns, we can see that inclusion of the realist variables is necessary whatever the effects of regime type and initiation. The posterior probabilities of the models without the realist variables, calculated from Equation (21), are approximately zero, across all values of φ. Of the models that include the realist variables, the posterior probabilities, calculated from Equation (21), are affected by the value of φ. As φ goes from 1 to 5, we see a shift from relatively strong support for model 4 to very strong support for model 3.[27] The explanation for the sensitivity is the same as it was for the dispute escalation example; models (3–5) are highly correlated.[28] The addition of the regime and

[26]In models 2 and 5, Reiter and Stam use a pair of fractional polynomials. The canonical correlation between models 1 and 2 and between models 4 and 5 are 1.

[27]When φ is equal to 1, the probability of model 4 is 67 percent. When φ is equal to 5, the probability of model 3 is 90 percent.

[28]The canonical correlations between these models are 1.

initiation variables makes little difference. If we look at the "central value" for φ (1.65), however, we see that model 4, which includes the regime type and initiation variables, is the best-supported model. There is some evidence then for Reiter and Stam's results, but that evidence is far from conclusive. There is little evidence for the fractional polynomial models.

Discussion

The substantive results of the analyses I have presented are surprising. There is a significant and steadily growing presumption, fueled by empirically based research such as that produced by Huth, Gelpi, and Bennett and Reiter and Stam, that realism is wrong or is insufficient to explain conflict outcomes in international relations. Some of the research upon which this presumption is based, however, is flawed. The analyses I address in this article contain comparative model tests that compare a realist model with a nonrealist model. These models are nonnested in the sense defined in an earlier section. The traditional methods of model comparison used by these authors are therefore inappropriate, and overstated conclusions have been the result.

The analyses I have presented challenge the conventional wisdom that realism performs poorly when compared to its rivals. Results that appeared certain in Huth, Gelpi, and Bennett now appear much less certain. While the Cox, Vuong, and Bayes results all lean slightly toward the rational deterrence model, the evidence is weak at best. The inconclusive results stem from the fact that the models are highly correlated. Clearing up the confusion would require respecifying one or both models in such a way as to diminish their correlation. Future research might also profitably examine the reasons why the models appear to be measuring the same underlying factors. An exploration of this curious result may lead to a synthesis of the two models that is more satisfactory than either of these models alone.

Reanalysis of the Reiter and Stam models also challenges the conventional wisdom regarding realism. The Reiter and Stam models, unlike the previous models, are not highly correlated, and the results of the three tests reflect this fact. In each test, the results robustly show that the realist model does a better job of accounting for war outcomes than the nonrealist model. In both cases, then, we see that realism, when appropriately tested against these particular rival theories, either does almost as well as the rival or better than the rival. Appropriate comparative tests are the key.

Figure 4 Comparing the Three Approaches

	Cox	Vuong	Bayes
Absolute	X		
Relative		X	X
Model Selection		X	X
Requires Simulation	X		
Hypothesis test	X	X	
Prior Information			X
Ease of calculation	2	1	3

On the methodological side, the results I have presented reinforce the idea that we must use statistical tests that are appropriate for the question being asked. Inappropriate tests produce misleading results. The Cox test, the Vuong test, and Bayes factors are appropriate techniques for discriminating between rival nonnested models. Which approach a nonmethodologist should use remains a question. All three techniques perform poorly when the models under consideration are highly correlated Clarke (1999). All three techniques perform well when the models under consideration are not highly correlated. The decision comes down to how easy the method is to perform and the degree to which the method is controversial.

Figure 4 summarizes the relative "pros and cons" of the three techniques. The Vuong test is the easiest to perform; it requires only calculation of the difference in the average log-likelihoods and calculation of the normalization. The Vuong test is also the least controversial of the three techniques. The test allows one to accept or reject a single null hypothesis much as one would with a t-test or an F-test. Furthermore, it requires neither simulation nor prior information. The Vuong test does have two disadvantages. First, the Vuong test is a "relative" as opposed to an "absolute" test and therefore cannot tell us whether or not *both* models under consideration are a bad fit for the data. Second, the Vuong test is a classical hypothesis test with all the concomitant problems of hypothesis tests.

The Cox test is harder to perform than the Vuong test. It is also more controversial than the Vuong test, but not as controversial as Bayes factors. Calculation of the Cox test, at least for probit and logit models, requires simulating the pseudomaximum-likelihood estimates that are used in the expected log-likelihood ratio. Extending this technique to different models is not difficult. The advantage of using the Cox test, the ability to reject both models, is also the test's most controversial aspect. Most tests allow a single decision to be made regarding a single null

hypothesis, which greatly eases the interpretative task. Opponents of the Cox test argue that if the test rejects both models, it is unclear how one should proceed.

Bayes factors based on the Laplace approximation are both the hardest to calculate and the most controversial. Calculation requires approximating the integrated likelihoods for both models. Software for this purpose has only been produced for generalized linear models with log, logit, or log-log link functions and either binomial or poisson errors.[29] Generalizing to other models is not straightforward. The technique is controversial for the same reason that all Bayesian techniques are controversial: it requires the specification of a prior distribution. The inclusion of prior information into empirical research has yet to be widely embraced by political methodologists and is still foreign to many substantive political scientists. The strength of the technique is that Bayes factors are not hypothesis tests and are therefore free of the controversies regarding sample size and p-values that surround hypothesis tests (see Morrison and Henkel 1970).

If only one of these techniques were to be used, I would argue for the Vuong test because of the relative ease of calculation and the fact that it is relatively uncontroversial. One can never go wrong, however, by using all three techniques simultaneously. The Vuong and Cox tests are complimentary in the sense that if the Cox rest rejects both models, the Vuong test can suggest a direction for further research. Bayes factors are useful for the simple reason that they are not hypothesis tests. If all three techniques reach similar substantive conclusions, as they do here, the results demonstrate a robustness that is absent when any of these techniques are used in isolation.

Directions for Further Research

The expository nature of this article leaves certain questions unanswered. With the exception of Bayes factors, the tests presented are paired tests. That is, they attempt to discriminate between two rival models. Researchers in international relations, however, are often faced with three or more models at a time. While work on joint versions of the Cox and Vuong tests for discrete choice models has begun (see McAleer 1995), more development is needed.

Empirical work in international relations is plagued with a lack of independence among units, across time, and across space. Without extensive Monte Carlo testing to assess the impact of this lack of independence on the methods presented here, my results, however compelling, must be considered preliminary.

Finally, I have not directly addressed the question of when the methods will give divergent answers. The question is a difficult one, as the three methods answer slightly different questions. The answer to this question will depend upon values such as the size of the sample and the correlation between the models. Preliminary Monte Carlo simulations have been suggestive (Clarke 1999), but again, more detailed analysis is necessary.

As I have demonstrated, the tasks of comparative theory testing and model selection in international relations, and throughout political science, often take place on an informal and *ad hoc* basis. Rarely are probabilistic statements made regarding rival models. Given the proliferation of theories and models in political science research, methods of discriminating between nonnested models should hold a more prominent place in the toolboxes of empirical scholars. This statement is as true for the American subfield as it is for international relations. Although there is more agreement on a general model in American politics, issues of functional form and error structure are still debated. With advances in statistical techniques and the concomitant advances in computing, there is no reason why probabilistic statements concerning rival models cannot be made. As I have demonstrated, such statements can make important contributions to prominent debates in the literature. Ideally, these methods will become as common and as familiar as the t-test.

Manuscript submitted January 13, 2000.
Final manuscript received December 1, 2000.

Appendix A
The Cox Test for Rival Probit Models

Consider the following univariate binary choice models[30]:

$$H_f: P\left(Y_t = 1 \middle| x_t\right) = \Phi\left(\theta' x_t\right) = \int_{-\infty}^{\theta' x_t} \phi(v)dv$$
$$H_g: P\left(Y_t = 1 \middle| z_t\right) = \Phi\left(\gamma' z_t\right) = \int_{-\infty}^{\gamma' x_t} \phi(v)dv$$

where $\Phi(\theta | x_t) = \Phi_1$ and $\Phi(\gamma | z_t) = \Phi_2$ are normal probability distribution functions and $\phi(.)$ is the density function of the standard normal variate.

[29] It is also possible, but not easy, to produce Bayes factors using MCMC methods (see Smith 1999; Clarke 2000).

[30] To calculate the Cox test for logit models, simply replace Φ with Λ in what follows.

The average log-likelihood functions under H_f and H_g are:

$$\begin{aligned} H_f: L_f\left(Y,\theta\middle|x\right) &= N^{-1}\sum_{t=1}^{N} Y_t \ln\Phi_{t1} + (1-Y_t)\ln\left(1-\Phi_{t1}\right) \\ H_g: L_g\left(Y,\gamma\middle|z\right) &= N^{-1}\sum_{t=1}^{N} Y_t \ln\Phi_{t2} + (1-Y_t)\ln\left(1-\Phi_{t2}\right) \end{aligned}.$$

The numerator of the Cox test statistic is:

$$T_f(R) = L_f\left(Y,\hat{\theta}\right) - L_g\left(Y,\hat{\gamma}\right) - C\left(\hat{\theta},\hat{\gamma}_*(R)\right)$$

where

$$C\left(\hat{\theta},\hat{\gamma}_*(R)\right) = \int \ln\left(\frac{f\left(y,\theta\right)}{g\left(y,\hat{\gamma}_*(R)\right)}\right) f\left(y,\hat{\theta}\right) dy.$$

For the binary choice models defined above:

$$L_f\left(Y,\hat{\theta}\right) - L_g\left(Y,\hat{\gamma}\right) = N^{-1}\sum_{t=1}^{N}\left[Y_t \ln\left(\frac{\hat{\Phi}_{t1}}{\hat{\Phi}_{t2}}\right) + \left(1-Y_t\right)\ln\left(\frac{1-\hat{\Phi}_{t1}}{1-\hat{\Phi}_{t2}}\right)\right]$$

and

$$C\left(\hat{\theta},\hat{\gamma}_*(R)\right) = N^{-1}\sum_{t=1}^{N}\left[Y_t \ln\left(\frac{\hat{\Phi}_{t1}}{\hat{\Phi}_{t2}(R)}\right) + \left(1-Y_t\right)\ln\left(\frac{1-\hat{\Phi}_{t1}}{1-\hat{\Phi}_{t2}(R)}\right)\right]$$

where $\hat{\Phi}_{t1} = \Phi(x_t'\hat{\theta})$ and $\hat{\Phi}^*_{t2}(R) = \Phi\left(z_t'\hat{\gamma}_*(R)\right)$.

Computing $\hat{\gamma}_*(R)$ by simulation:

- Artificially simulate y using $f(y,\hat{\theta})$ as the DGP.
- $Y_j = (Y_{1j}, Y_{2j}, \ldots, Y_{nj})'$ are the independent observations generated artificially according to $f(y,\hat{\theta})$.
- Use Y_j to compute the maximum likelihood estimate of γ under $f(y,\hat{\theta})$ and call it $\hat{\gamma}_j$.
- Then estimate γ_* by

$$\hat{\gamma}_*(R) = \frac{1}{R}\sum_{j=1}^{R}\hat{\gamma}_j$$

where R stands for the number of repetitions.

The variance of the Cox statistic is:

$$N^{-1}d'\left\{I_N - R(\hat{\beta})\left[R'(\hat{\beta})R(\hat{\beta})\right]^{-1}R'(\hat{\beta})\right\}d$$

where

$$d = N^{-1}\sum_{t=1}^{N}\left[Y_t \ln\left(\frac{\hat{\Phi}_{t1}}{\hat{\Phi}_{t2}}\right) + \left(1-Y_t\right)\ln\left(\frac{1-\hat{\Phi}_{t1}}{1-\hat{\Phi}_{t2}}\right)\right]$$

and

$$R(\beta) = \begin{bmatrix} 1 & \frac{\partial \ln f\left(Y_1,\beta\right)}{\partial\beta_1} & \cdots & \frac{\partial \ln f\left(Y_1,\beta\right)}{\partial\beta_p} \\ 1 & \frac{\partial \ln f\left(Y_2,\beta\right)}{\partial\beta_1} & \cdots & \frac{\partial \ln f\left(Y_2,\beta\right)}{\partial\beta_p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \frac{\partial \ln f\left(Y_N,\beta\right)}{\partial\beta_1} & \cdots & \frac{\partial \ln f\left(Y_N,\beta\right)}{\partial\beta_p} \end{bmatrix}$$

Appendix B
The Vuong Test for Rival Logit Models

Consider the following univariate binary choice models[31]:

$$\begin{aligned} H_f: P\left(Y_t = 1\middle|x_t\right) &= \Lambda\left(\theta' x_t\right) = \frac{e^{\theta' x_t}}{1+e^{\theta' x_t}} \\ H_g: P\left(Y_t = 1\middle|z_t\right) &= \Lambda\left(\gamma' z_t\right) = \frac{e^{\gamma' z_t}}{1+e^{\gamma' z_t}} \end{aligned}$$

where $\Lambda(\theta|\, x_t) = \Lambda_1$ and $\Lambda(\gamma|\, z_t) = \Lambda_2$ are logistic probability distribution functions.

The average log-likelihood functions under H_f and H_g are:

$$\begin{aligned} H_f: L_f\left(Y,\theta\middle|x\right) &= N^{-1}\sum_{t=1}^{N} Y_t \ln\Lambda_{t1} + \left(1-Y_t\right)\ln\left(1-\Lambda_{t1}\right) \\ H_g: L_g\left(Y,\gamma\middle|z\right) &= N^{-1}\sum_{t=1}^{N} Y_t \ln\Lambda_{t2} + \left(1-Y_t\right)\ln\left(1-\Lambda_{t2}\right) \end{aligned}.$$

The Vuong test statistic with the BIC correction is:

$$\frac{L\tilde{R}_n\left(\hat{\theta}_n,\hat{\gamma}_n\right)}{\left(\sqrt{n}\right)\hat{\omega}}$$

where

$$L\tilde{R}_n\left(\hat{\theta}_n,\hat{\gamma}_n\right) \equiv LR_n\left(\hat{\theta}_n,\hat{\gamma}_n\right) - K_n\left(F_\theta, G_\gamma\right)$$

and

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

[31] To calculate the Vuong test for probit models, simply replace Λ with Φ in what follows.

For the binary choice models defined above:

$$L\tilde{R}\left(\hat{\theta}_n, \hat{\gamma}_n\right) = N^{-1}\sum_{t=1}^{N}\left[Y_t \ln\left(\frac{\hat{\Lambda}_{t1}}{\hat{\Lambda}_{t2}}\right) + \left(1-Y_t\right)\ln\left(\frac{1-\hat{\Lambda}_{t1}}{1-\hat{\Lambda}_{t2}}\right)\right] - \left[\left(\frac{p}{2}\right)\ln n - \left(\frac{q}{2}\right)\ln n\right]$$

and

$$\hat{\omega}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left[Y_t \ln\left(\frac{\hat{\Lambda}_{t1}}{\hat{\Lambda}_{t2}}\right) + \left(1-Y_t\right)\ln\left(\frac{1-\hat{\Lambda}_{t1}}{1-\hat{\Lambda}_{t2}}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{\hat{\Lambda}_{t1}}{\hat{\Lambda}_{t2}}\right) + \left(1-Y_t\right)\ln\left(\frac{1-\hat{\Lambda}_{t1}}{1-\hat{\Lambda}_{t2}}\right)\right]^2.$$

References

Achen, Christopher H. 1982. *Interpreting and Using Regression.* Beverly Hills: Sage.

Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Proceedings of the Second International Symposium on Information Theory,* ed. B. N. Petrov and F. Csaki. Budapest: Akademinai Kiado.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:641–674.

Bennett, D. Scott. 1997. "Testing Alternative Models of Alliance Duration, 1816-1984." *American Journal of Political Science* 41:846–878.

Bennett, D. Scott, and Allan C. Stam. 1996. "The Duration of Interstate Wars, 1816-1985." *American Political Science Review* 90:239–257.

Benoit, Kenneth. 1996. "Democracies Really Are More Pacific (in General): Reexamining Regime Type and War Involvement." *Journal of Conflict Resolution* 40:636–657.

Clarke, Kevin A. 1999. "Nonnested Tests and the Escalation of Great Power Militarized Disputes." Unpublished manuscript. University of Michigan.

Clarke, Kevin A. 2000. "The Effect of Priors on Approximate Bayes Factors from MCMC Output." Unpublished manuscript. University of Michigan.

Cox, David R. 1961. "Tests of Separate Families of Hypotheses." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* vol. I. Berkeley: University of California Press.

Cox, David R. 1962. "Further Results on Tests of Separate Families of Hypotheses." *Journal of the Royal Statistical Society Series B* 24:406–424.

Davenport, Christian. 1999. "Human Rights and the Democratic Proposition." *Journal of Conflict Resolution* 43:92–116.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Enterline, Andrew J. 1998. "Regime Changes, Neighborhoods, and Interstate Conflict, 1816-1992." *Journal of Conflict Resolution* 42:804–829.

Feng, Yi, and Paul J. Zak. 1999. "The Determinants of Democratic Transitions." *Journal of Conflict Resolution* 43:162–177.

Gelpi, Christopher. 1997. "Democratic Diversions: Governmental Structure and the Externalization of Domestic Conflict." *Journal of Conflict Resolution* 41:255–282.

Gourieroux, Christian, and Alain Monfort. 1995. *Statistics and Econometric Models,* vol. 2. New York: Cambridge University Press.

Granger, C.W.J., Maxwell L. King, and Halbert White. 1995. "Comments on Testing Economic Theories and the Use of Model Selection Criteria." *Journal of Econometrics* 67:173–187.

Greene, William H. 1997. *Econometric Analysis.* 3rd ed. Upper Saddle River, N.J.: Prentice-Hall.

Hermann, Margaret G., and Charles W. Kegley. 1996. "Ballots, a Barrier Against the Use of Bullets and Bombs: Democratization and Military Intervention." *Journal of Conflict Resolution* 40:436–460.

Horowitz, Joel L. 1983. "Statistical Comparison of Non-Nested Probabilistic Discrete Choice Models." *Transportation Science* 17:319–350.

Howson, Colin, and Peter Urbach. 1993. *Scientific Reasoning: The Bayesian Approach.* 2nd ed. Chicago: Open Court.

Huth, Paul, D. Scott Bennett, and Christopher Gelpi. 1992. "System Uncertainty, Risk Propensity, and International Conflict Among the Great Powers." *Journal of Conflict Resolution* 36:478–517.

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* 87:609–623.

Huth, Paul, and Bruce Russett. 1993. "General Deterrence Between Enduring Rivals: Testing Three Competing Models." *American Political Science Review* 87:61–73.

Jeffreys, Harold. 1961. *Theory of Probability.* 3rd ed. Oxford: Clarendon.

Johnson, Richard A., and Dean W. Wichern. 1998. *Applied Multivariate Statistical Analysis.* 4th ed. Upper Saddle River, N.J.: Prentice-Hall.

Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee. 1985. *The Theory and Practice of Econometrics.* 2nd ed. New York: Wiley.

Kennedy, Peter. 1992. *A Guide to Econometrics.* 3rd ed. Cambridge: The MIT Press.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30:666–687.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* New York: Cambridge University Press.

Kmenta, Jan. 1986. *Elements of Econometrics.* 2nd ed. New York: Macmillan.

Kullback, Solomon. 1959. *Information Theory and Statistics.* New York: Wiley.

Lai, Brian, and Dan Reiter. 2000. "Democracy, Political Similarity, and International Alliances, 1816-1992." *Journal of Conflict Resolution* 44:203–227.

Lemke, Douglas, and William Reed. 1996. "Regime Types and Status Quo Evaluations: Power Transition Theory and the Democratic Peace." *International Interactions* 22:143–164.

Maoz, Zeev, and Bruce Russett. 1993. "Normative and Structural Causes of Democratic Peace, 1946-1986." *American Political Science Review* 87:624–638.

McAleer, Michael. 1987. "Specification Tests for Separate Models: A Survey." In *Specification Analysis in the Linear Model*, ed. M.L. King and D.E.A. Giles. London: Routledge and Kegan Paul.

McAleer, Michael. 1995. "The Significance of Testing Empirical Non-Nested Models." *Journal of Econometrics* 67:149–171.

Morrison, Denton E., and Ramon E. Henkel (eds.). 1970. *The Significance Test Controversy: A Reader.* Chicago: Aldine.

Morrow, James D., Randolph M. Siverson, and Tressa E. Taberes. 1998. "The Political Determinants of International Trade: The Major Powers, 1907-90." *American Political Science Review* 92:649–661.

Palmer, Glenn, and J. Sky David. 1999. "Multiple Goals or Deterrence: A Test of Two Models in Nuclear and Nonnuclear Alliances." *Journal of Conflict Resolution* 43:748–770.

Pesaran, M.H. 1987. "Global and Partial Non-Nested Hypotheses and Asymptotic Local Power." *Econometric Theory* 3:69–97.

Pesaran, M.H., and B. Pesaran. 1993. "A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Nonnested Models." *Journal of Econometrics* 57:377–392.

Pollins, Brian M. 1996. "Global Political Order, Economic Change, and Armed Conflict: Coevolving Systems and the Use of Force." *American Political Science Review* 90:103–117.

Raftery, Adrian E. 1993. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models." Tech. Rep. 255, University of Washington.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research (with Discussion)." In *Sociological Methodology 1995,* ed. P.V. Marsden. Cambridge: Blackwell.

Rasler, Karen, and William R. Thompson. 1999. "Predatory Initiators and Changing Landscapes for Warfare." *Journal of Conflict Resolution* 43:411–433.

Reiter, Dan, and Allan C. Stam. 1998a. "Democracy and Battlefield Military Effectiveness." *Journal of Conflict Resolution* 42:259–277.

Reiter, Dan, and Allan C. Stam. 1998b. "Democracy, War Initiation, and Victory." *American Political Science Review* 92:377–389.

Schelling, Thomas. 1960. *The Strategy of Conflict.* Cambridge: Harvard University Press.

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

Shin, Michael, and Michael D. Ward. 1999. "Lost in Space: Political Geography and the Defense-Growth Trade-Off." *Journal of Conflict Resolution* 43:793–817.

Signorino, Curtis S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93:279–298.

Signorino, Curtis S., and Kuzey Yilmaz. 2000. "Strategic Misspecification in Discrete Choice Models." Unpublished manuscript. University of Rochester.

Smith, Alastair. 1996. "To Intervene or Not to Intervene: A Biased Decision." *Journal of Conflict Resolution* 40:16–40.

Smith, Alastair. 1999. "Testing Theories of Strategic Choice: The Example of Crisis Escalation." *American Journal of Political Science* 43:1254–1283.

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

Waltz, Kenneth N. 1979. *Theory of International Politics.* Reading, Mass.: Addison-Wesley.

White, Halbert. 1994. *Estimation, Inference and Specification Analysis.* New York: Cambridge University Press.