# Discriminating Methods: Tests for Nonnested Discrete Choice Models[§]

Kevin A. Clarke[†] Curtis S. Signorino[‡]

**Abstract**

We consider the problem of choosing between rival statistical models that are nonnested in terms of their functional forms. We assess the ability of two tests, one parametric and one distribution-free, to discriminate between such models. Our monte carlo simulations demonstrate that both tests are, to varying degrees, able to discriminate between strategic and nonstrategic discrete choice models. The distribution-free test appears to have greater relative power in small samples. As an application, we consider the process of European Union enlargement and demonstrate that a strategic model captures this process better than a standard selection model.

September 13, 2006

[§]Earlier versions of this paper were presented at the 2003 annual meetings of the International Studies Association, the Society for Political Methodology, and the American Political Science Association; we thank the participants for their comments. Support from the National Science Foundation (Grant SES-0213771) is gratefully acknowledged.

[†]Corresponding author. Assistant Professor, Department of Political Science, University of Rochester, `kevin.clarke@rochester.edu`.

[‡]Associate Professor, Department of Political Science, University of Rochester, `curt.signorino@rochester.edu`.

# 1 Introduction

The empirical study of political science has, in the last ten years, undergone something akin to a sea-change. Where it was once common for political scientists to employ a linear functional form regardless of the theory being tested, we now see new attention being paid to the connection between theory and model (Morton 1999, Signorino 1999, Signorino and Yilmaz 2003). The result of this attention has been an expansion in the number of different functional forms being employed by quantitative political scientists.

This increase in the number of modelling choices available to researchers has brought with it new challenges. For example, although Signorino (1999) demonstrates that traditional specifications of statistical models are generally inconsistent with strategic theories of political science, no rigorous framework has emerged for comparing strategic models against one another, or against nonstrategic models. While it is clear that strategic specifications provide different answers from traditional specifications, it not yet clear that these strategic specifications are, in fact, superior. We therefore need a procedure to determine whether one specification is "closer" than another specification to the data generating process (DGP).

A related problem stems from the fact that not all theory is detailed enough to allow the derivation of a functional form suitable for testing. An empirical researcher faced with choosing a statistical specification under these conditions needs guidance in choosing between the many functional forms, some strategic and some not, that may be used to model political phenomena.

In either of these cases, empirical researchers need tools that allow comparisons to be made between models with different functional forms. Such models, however, are generally nonnested (neither model is a special case of the other model).[1] Discriminating between nonnested models requires specialized tests that are rarely used in political science research. Clarke (2001) introduced the issue of nonnested testing to political science, and Clarke (2003a) introduced a simple distribution-free test for nonnested model discrimination. These articles, however, consider only models that are nonnested in terms of their covariates. Testing models that are nonnested in terms of their functional forms is a natural extension of this line of research.

The goals of this article are threefold. First, we briefly review the role of theory in determining the appropriate functional form of a discrete choice model. Second, we briefly review two non-nested model discrimination tests and demonstrate that discriminating between discrete choice models with different functional forms is possible, even with small samples. Third, we provide, with the help of an example from the literature, specific advice on how to determine whether or not a strategic specification fits the data better than a traditional specification.

The article proceeds as follows. In the next section, we present a common crisis scenario and consider three functional forms a researcher might choose

[1] See Clarke (2001) for a technical definition of "nonnested."

when modelling it: a probit model, a selection model, and a strategic model. The natural question is the extent to which we can discriminate between these models, given that the data are generated by a strategic process. We answer this question by conducting a monte carlo experiment that assesses the relative power of the Vuong and distribution-free tests. We find that both the Vuong and distribution-free tests are able to discriminate between the models accurately. In large samples, the two tests are essentially equivalent. In small samples, however, the distribution-free test outperforms the Vuong test. Finally, we discuss an empirical example concerning European Union enlargement and provide specific advice for the researcher.

## 2 Competing Discrete Choice Models

Consider a researcher who wants to model the conditions under which two states are likely to go to war. Figure 1 displays a simple conflict scenario, which we use throughout the paper. In this crisis situation, state 1 must decide whether to attack ($A$) state 2 or not attack ($\neg A$). If attacked, state 2 then chooses whether to resist ($R$) or not resist ($\neg R$). If state 1 does not attack, we assume that the status quo ($SQ$) is maintained. If state 1 attacks and state 2 backs down, we assume that state 2 capitulates ($Cap_2$). Finally, if the state 1 attacks and the state 2 resists, we assume that war ($War$) is

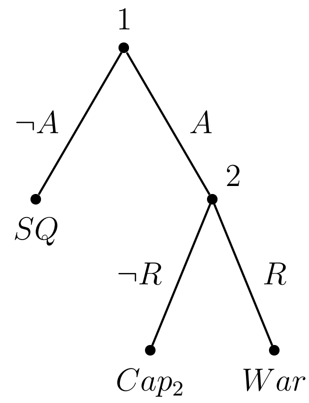



Figure 1: Conflict Model. *State 1 decides whether to attack (A) state 2 or not attack (¬A). If attacked, state 2 decides whether to resist (R) or not resist (¬R). The states' actions lead to three outcomes: the status quo (SQ), capitulation by state 2 ($Cap_2$), or war ($War$).*

the result.[2]

Although we have already imposed certain constraints on the conflict scenario, there remains a great deal of latitude for researchers who wish to model it. A simple point, but one that is often overlooked, is that the statistical model a researcher employs depends on his or her underlying theory of the process generating the data. Some researchers take an explicitly game-theoretic approach and derive their functional form directly from their model. Others rely on the structure of the available data; researchers with binary data tend to use different statistical models than those researchers with sequential data. Given rival models that represent two different data generating processes, we need to be able to test which model is better supported by the data.

To make this point more concretely, we turn to three different models a re-

[2]The scenario could just as well represent an extended deterrence situation where state 1 is threatening to attack a protege of state 2. If the protege is attacked, state 2 must decide whether to defend its protege (see Huth 1988, Signorino and Tarar 2002).

searcher might choose when empirically analyzing the conflict scenario in Figure 1. We highlight these models as they appear throughout the international relations literature. We begin with a probit model, follow with a selection model, and end with the two variants (binary and sequential) of a strategic model.

## 2.1 Probit Model

Due to the widespread availability of binary data — and the commensurate dearth of sequential data — the most popular method of analyzing a conflict scenario such as in Figure 1 has been the probit or logit model.[3] Given this modelling choice, two of the outcomes in the conflict scenario, status quo ($SQ$) and capitulation by state 2 ($Cap_2$), are usually aggregated into a single outcome, the absence of war ($\neg War$).

Figure 2(a) provides graphical intuition about the data and the model. States 1 and 2 either go to war or not. The propensity to go to war, $y^*_{War}$, is a linear function of a set of regressors pertaining to state 1 and of a set of regressors pertaining to state 2,

$$y^*_{War} = X\beta + Z\gamma + \epsilon.$$

$X$ in the above equation represents state 1's regressors, $Z$ represents state 2's regressors, $\beta$ and $\gamma$ are coefficient vectors on $X$ and $Z$, respectively, and

[3] We focus solely on the probit model as logit and probit are almost perfect substitutes for one another.

$\epsilon$ is a random disturbance, assumed to be normally distributed with mean zero and variance one.[4]

We do not observe the latent variable, $y^*_{War}$, but only whether the joint propensity is above or below the threshold for war,

$$y_{War} = \begin{cases} 1, & \text{if } y^*_{War} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Maximum likelihood estimation of the probit model is based on the resulting probabilities,

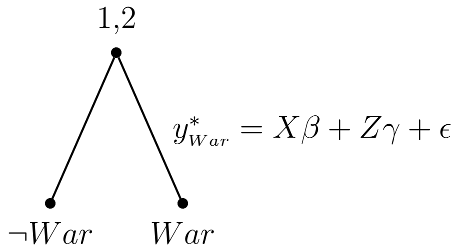
$$\Pr(y_{War} = 0|X, Z) = 1 - \Phi(X\beta + Z\gamma) \quad (1)$$

$$\Pr(y_{War} = 1|X, Z) = \Phi(X\beta + Z\gamma) \quad (2)$$

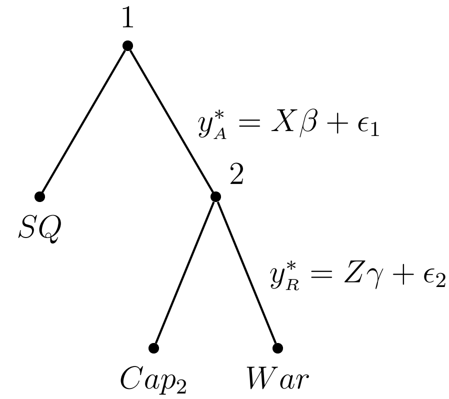
where $\Phi(\cdot)$ is the distribution function of the standard normal.

Throughout the remainder of the paper, we motivate the statistical models by making random utility assumptions, as opposed to relying on the structure of the data. It is important to note that probit models may also be motivated by random utility assumptions (Judge, Griffiths, Hill, Lutkepohl, and Lee 1985). In random utility models, decision-makers are assumed to have preferences over outcomes, which are represented by their utilities for those outcomes.

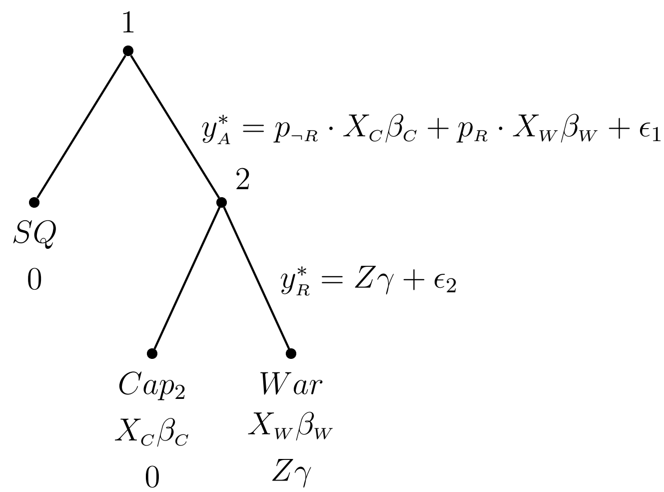
[4] This last assumption is made in order to identify the model and is innocuous (Greene 2003).




(a) Probit Model




(b) Selection Model




(c) Strategic Model

Figure 2: Alternative Discrete Choice Specifications. *Figure (a) shows the common probit specification. In the selection model, Figure (b), the* $War$ *outcome results from a "selection" equation,* $y^*_A$*, and an "outcome" equation,* $y^*_R$*, with the additional assumption that* $\epsilon_1$ *and* $\epsilon_2$ *are correlated. Finally, Figure (c) displays a strategic model, with each player's payoffs shown below the outcomes. In this case,* $War$ *is also a result of state 1's and 2's decisions. However, state 1's decision,* $y^*_A$*, is based on an expected utility calculation, and the disturbances are assumed to be uncorrelated.*

Each decision-maker chooses the option available to her for which she has the highest utility. Because the empirical analyst does not fully observe the decision-makers' utilities, the analyst models each player's utility as having an observable component and a random component.[5]

It is doubtful that the probit version of the conflict scenario could be reasonably motivated by random utility assumptions. The analyst would need to make one of two assumptions: either (1) that states 1 and 2 jointly make a decision between $War$ and $\neg War$; or, (2) that their individual actions somehow lead to either $War$ or $\neg War$. Both assumptions have obvious theoretical problems, since both "black box" important components of the decision-making process. Under the first assumption, we must assume the existence of an unobserved decision aggregation rule that is consistent with considering the dyad as a single decision-making unit. Under the second assumption, we must assume an unobserved sequence of choices that are consistent with considering only the $War$ versus $\neg War$ outcomes. With that said, we turn to two models that can be more directly motivated by random utility assumptions.

## 2.2 Selection Model

Now consider a researcher who has sequential data on state 1's decision to attack and state 2's decision to resist after being attacked. One random

[5] For examples of strategic and nonstrategic random utility models, see Signorino (2002).

utility-based modelling option available to the researcher is the Heckman selection model, which has become increasingly popular in economics and in political science.[6]

The selection model displayed in Figure 2(b) retains the original sequential choice structure depicted in Figure 1. State 1 decides whether to attack or not based on a comparison of its utility for attacking, $U_1(A)$, with its utility for not attacking, $U_1(\neg A)$. The "selection equation,"

$$y_A^* = X\beta + \epsilon_1 \tag{3}$$

represents state 1's net utility for attacking, where $X\beta$ is the observable component of its utility and $\epsilon_1$ is the random term.[7] As a utility maximizer, state 1 attacks when $y_A^* > 0$. We observe,

$$y_A = \begin{cases} 1, & \text{if } y_A^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

If state 1 attacks (i.e., $y_A = 1$), then state 2 must decide whether to resist or not. Again, this decision is based on a comparison of state 2's utility for going to war versus capitulating. The "outcome equation,"

$$y_R^* = Z\gamma + \epsilon_2,$$

[6]See Greene (2003), Reed (2000), and Signorino (2002).

[7]Equation 3 represents the difference in state 1's utilities for attacking versus not attacking. Equivalently, if we normalize state 1's utility for the status quo to zero, then it represents its utility for attacking.

represents state 2's net utility for resisting, where $Z\gamma$ is the observable component of its utility and $\epsilon_2$ is the random term. As a utilty maximizer, state 2 resists whenever $y_R^* > 0$. We observe,

$$y_R = \begin{cases} 1, & \text{if } y_R^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The final step in specifying the selection model concerns the disturbances, $\epsilon_1$ and $\epsilon_2$. Following Greene (2003), we assume the disturbances are distributed bivariate normal with mean zero, variance one, and correlation $\rho$. Given these assumptions, the probability of observing each outcome is,

$$\begin{aligned} \Pr(SQ) = \Pr(y_A = 0|X, Z) &= 1 - \Phi(X\boldsymbol{\beta}) \\ \Pr(Cap_2) = \Pr(y_A = 1, y_R = 0|X, Z) &= \Phi_2(X\boldsymbol{\beta}, -Z\gamma, -\rho) \\ \Pr(War) = \Pr(y_A = 1, y_R = 1|X, Z) &= \Phi_2(X\boldsymbol{\beta}, Z\gamma, \rho), \end{aligned}$$

where $\Phi(\cdot)$ and $\Phi_2(\cdot)$ are the c.d.f.'s of the standard normal and standard bivariate normal, respectively, and $\rho$ is the correlation between the error terms for the two equations.

## 2.3 Strategic Choice Model

A third modelling choice available to the researcher analyzing the conflict scenario is to assume that the choices both states make occur not only sequentially, but strategically. For example, in the selection model discussed in the previous section, state 1's decision to attack or not (Equation 3) is a linear function and does not take into account what state 2 is likely to do. In contrast, assume that state 1 chooses between the status quo and attacking state 2 taking into consideration whether it believes that state 2 will capitulate or choose war. Given that state 1 chooses to attack, state 2 then decides between capitulation and war based on a straightforward utility maximization.

Figure 2(c) displays such a strategic choice model of the crisis scenario. As in Figure 1, we assume that state 1 attacks ($A$), or does not attack ($\neg A$). If attacked, state 2 must decide whether to resist ($R$) or not resist ($\neg R$). The payoffs to each state are given below the outcomes in Figure 2(c). We normalize the status quo payoff for state 1 to zero. Whereas in the previous models we combined the factors that influenced state 1's decision into $X\beta$, we now separate them into (1) those that affect state 1's payoff for the capitulation outcome ($X_C\beta_C$) and (2) those that affect state 1's payoff for the war outcome ($X_W\beta_W$). As before, we normalize state 2's payoff for the status quo at zero, and we let its payoff for war be $Z\gamma$. We assume that a disturbance is associated with the expected utilities at each information set, and that the disturbances are independently distributed standard normal.[8]

To derive the strategic probability model, we work "up the game," starting with state 2's decision. If attacked, state 2 considers only whether to resist or not. As in the selection model,

$$y_R^* = Z\gamma + \epsilon_2,$$

represents state 2's net utility for resisting. $Z\gamma$ is the observable component of the utility, and $\epsilon_2$ is the random term. As a utility maximizer, state 2 resists whenever $y_R^* > 0$. We observe,

$$y_R = \begin{cases} 1, & \text{if } y_R^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Given the distributional assumption for $\epsilon_2$, state 2's choice probabilities are,

$$\begin{aligned} p_R = \Pr(y_R = 1|X, Z) &= \Phi(Z\gamma) \\ p_{\neg R} = \Pr(y_R = 0|X, Z) &= 1 - \Phi(Z\gamma). \end{aligned}$$

Now consider state 1's decision. As before, state 1's decision whether to attack is based on a comparison of its utility for attacking versus its utility for the status quo. In contrast to the selection model, however, we now assume that state 1 conditions its behavior on what it expects state 2 to do.

[8] This is the probit agent error model discussed in more detail in Signorino (2002). For a logit version, see McKelvey and Palfrey (1998) and Signorino (1999).

Because state 1 does not perfectly observe state 2's utilities, state 1 can only estimate the probability that state 2 will resist or not. Therefore, state 1's utility for attacking is an expected utility, based on the lottery representing whether state 2 will resist or not.

Since we normalize state 1's utility for the status quo to zero,

$$y_A^* = p_{\neg R} \cdot X_C \beta_C + p_R \cdot X_W \beta_W + \epsilon_1, \tag{4}$$

represents state 1's net expected utility for attacking. $p_{\neg R} \cdot X_C \beta_C + p_R \cdot X_W \beta_W$ is the observable component of the expected utility, and $\epsilon_1$ is the random utility component. State 1 attack when $y_A^* > 0$. We observe,

$$y_A = \begin{cases} 1, & \text{if } y_A^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

State 1's equilibrium choice probabilities are then,

$$\begin{aligned} p_A = \Pr(y_A = 1 | X, Z) &= \Phi(p_{\neg R} \cdot X_C \beta_C + p_R \cdot X_W \beta_W) \\ p_{\neg A} = \Pr(y_A = 0 | X, Z) &= 1 - \Phi(p_{\neg R} \cdot X_C \beta_C + p_R \cdot X_W \beta_W). \end{aligned}$$

Because the disturbances are independently distributed, the equilibrium outcome probabilities are the product of the choice probabilities along the path,

$$\Pr(SQ) = \Pr(y_A = 0|X, Z) \quad = \quad p_{\neg A} \tag{5}$$

$$\Pr(Cap_2) = \Pr(y_A = 1, y_R = 0|X, Z) \quad = \quad p_A \cdot p_{\neg R} \tag{6}$$

$$\Pr(War) = \Pr(y_A = 1, y_R = 1|X, Z) \quad = \quad p_A \cdot p_R. \tag{7}$$

Maximum likelihood estimation of the effect parameters, $\beta_C$, $\beta_W$, and $\gamma$, is based on these equilibrium probabilities assuming that the dependent variable denotes which of the three outcomes occurred for each observation.

### 2.3.1 A Binary Data Version of the Strategic Choice Model

The two discrimination tests discussed in the next section require that both rival models have precisely the same dependent variable. This requirement is problematic for researchers who wish to discriminate between the ubiquitous probit model and the more recent strategic model. The reason is that the strategic model has three outcomes, and the probit model has only two. The problem is easily solved, however. For a valid comparison of the two models, we simply need a version of the strategic model that has been aggregated for binary data.

Recall from Figures 1 and 2(a) that the binary data represent $War$ versus $\neg War$. Where the probit model ignores the different outcomes that comprise $\neg War$, the strategic model forces us to confront them. Thus, the

$\neg War$ outcome, in the binary version of the strategic model, is equivalent to the occurrence of either the status quo ($SQ$) *or* capitulation ($Cap_2$). The probability of $\neg War$ is therefore the probability of the status quo plus the probability of capitulation. The probabilities for the binary data version of the strategic model are then,

$$\Pr(\neg War) = p_{\neg A} + p_A \cdot p_{\neg R} \tag{8}$$

$$\Pr(War) = p_A \cdot p_R, \tag{9}$$

where the choice probabilities are those previously derived for the full strategic model. The $War$ outcome is the same in both versions of the model, therefore, the probability of war is the same in both models. These probabilities form the basis for maximum likelihood estimation of the effect parameters given binary data.

# 3 Nonnested Model Testing

The models in Section 2 are nonnested in terms of their functional forms.[9] Determining which of these functional forms is closest to the true, but unknown, specification requires the use of discrimination tests that are still new to the vast majority of political scientists. Two of the easiest and least contro-

[9]See Clarke (2001) for a definition of nonnested and methods of determining whether rival models are nonnested.

versial of these tests are the Vuong test (Vuong 1989) and a distribution-free test introduced by Clarke (2003a).

Both tests are based on the Kullback-Leibler information criteria (Kullback and Leibler 1951), which is simply a measure of similarity between an estimated model and the true distribution. We can write the Kullback-Liebler information criteria (KLIC) heuristically as

$$\text{KLIC} \equiv \text{True distribution} - \text{Statistical model.}$$

The point of the above equation is that we wish to choose the statistical model that minimizes the KLIC because that is the statistical model that is closest to the truth. Written more formally, the KLIC is

$$\text{KLIC} \equiv E_0[\ln h(\mathbf{Y}|\mathbf{X})] - E_0[\ln f(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})],$$

where $h$ is the true conditional density of $\mathbf{Y}$ given $\mathbf{X}$ (the true, but unknown distribution), $f$ is the statistical model, and $E_0$ is the expectation under the true distribution. Given that we wish to minimize the KLIC, we want to choose the model with the largest log-likelihoods, $E_0[\ln f(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})]$. In other words, we should choose one statistical model over another statistical model if the individual log-likelihoods of that model are significantly larger than the individual log-likelihoods of the rival model. The difference between the two tests we discuss here is whether we consider the average difference in the log-likelihoods of two models or the median difference in the log-likelihoods of two models.

## 3.1 The Vuong Test

The Vuong test considers the average difference in the log-likelihoods of two competing statistical models. The null hypothesis of the test is that this average difference is zero. Letting $f$ denote model 1, which has covariates $\mathbf{X}$ and coefficients $\boldsymbol{\beta}$, and $g$ denote model 2, which has covariates $\mathbf{Z}$ and coefficients $\boldsymbol{\gamma}$, we can write the null as

$$H_0 : E_0 \left[ \ln \frac{f(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})}{g(\mathbf{Y}|\mathbf{Z};\boldsymbol{\gamma})} \right] = 0.$$

The null hypothesis simply states that the two models are equally close to the true specification.

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

$$\frac{1}{n} LR(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \xrightarrow{a.s.} E_0 \left[ \ln \frac{f(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})}{g(\mathbf{Y}|\mathbf{Z};\boldsymbol{\gamma})} \right],$$

where $LR(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is the estimated difference in the log-likelihoods of the two models, or $L_f(\hat{\boldsymbol{\beta}}) - L_g(\hat{\boldsymbol{\gamma}})$. Thus, the expected value can be consistently estimated by $\left(\frac{1}{n}\right)$ times the usual log-likelihood ratio statistic.

Suitably normalized, the test statistic is normally distributed under the null hypothesis,

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

where the estimated variance is computed in the usual way (sum of the squares minus square of the sums),

$$\hat{\omega}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{f(\mathbf{Y}|\mathbf{X};\hat{\boldsymbol{\beta}})}{g(\mathbf{Y}|\mathbf{Z};\hat{\boldsymbol{\gamma}})}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(\mathbf{Y}|\mathbf{X};\hat{\boldsymbol{\beta}})}{g(\mathbf{Y}|\mathbf{Z};\hat{\boldsymbol{\gamma}})}\right]^2.$$

Thus, the 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.

The log-likelihoods used in the Vuong test are 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) information criteria or Schwarz's (1978) Bayesian information criteria. In the simulations that follow, we use the latter, making the adjusted statistic[10],

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

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

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

### 3.2 The Distribution-Free Test

Where the Vuong test considers the average difference in the log-likelihoods of two competing statistical models, the distribution-free test considers the median difference in the log-likelihoods of two competing statistical models (Clarke 2003a). Thus, if the models are equally close to the truth, half of the individual 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 of the individual log-likelihood ratios should be greater than zero. Conversely, if model $g$ is "better" than model $f$, more than half of the individual log-likelihood ratios should be less than zero.

The null hypothesis of the distribution-free test, therefore, is that if two models are equally close to the truth, half of the log-likelihood ratios should be great than zero,

$$H_0 : \Pr\left[\ln \frac{f(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})}{g(\mathbf{Y}|\mathbf{Z};\boldsymbol{\gamma})} > 0\right] = 0.5.$$

Letting $d_i = \ln f(\mathbf{Y}|\mathbf{X};\hat{\boldsymbol{\beta}}) - g(\mathbf{Y}|\mathbf{Z};\hat{\boldsymbol{\gamma}})$, the test statistic is

$$B = \sum_{i=1}^{n} I_{(0,+\infty)}(d_i),$$

where $I$ is the indicator function. The test statistic, therefore, is simply the number of positive differences, and it is distributed Binomial with parameters $n$ and $\theta = 0.5$.[11]

One of the great strengths of this test is that implementation is remarkably simple; the test can be produced by any mainstream statistical software package using the following algorithm[12]:

1. Run model $f$, saving the individual log-likelihoods,
2. Run model $g$, saving the individual log-likelihoods,
3. Compute the differences, $d_i$, and count the number of positive values, $B$,
4. The number of positive differences, $B$, is distributed binomial$(n, 0.5)$.

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

[11] The assumptions of the distribution-free test are unsurprising and general: the log-likelihood ratios are mutually independent, and each ratio is drawn from a continuous population (not necessarily the same) that has a common median $\theta$.

[12] In what follows, steps 1 and 2 are in the process of being implemented by STATA. Step 3 already exists in STATA because we are 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.

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

where $p$ and $q$ are the number of estimated coefficients in models $f$ and $g$, respectively. As we are working with the individual log-likelihood ratios, we cannot apply this correction to the "summed" log-likelihood ratio as we did for the Vuong test. We can, however, apply the *average* correction to the individual log-likelihood ratios. That is, we 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:

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

A discussion of the efficiency of the distribution-free test versus the Vuong test as well as proofs of the consistency and unbiasedness of the distribution-free test are found in Clarke (2003b).

# 4 Monte Carlo Simulations

We wish to determine if we can discriminate (1) between the strategic model and the selection model, and (2) between the binary data version of the strategic model and the probit model. To that end, we perform a suite of monte carlo simulations. In addition to answering our main question, the results also indicate under what conditions we can expect either the Vuong test or the distribution-free test to have greater relative power.

## 4.1 Experimental Design

The data generating process (DGP) for the experiment is the strategic model. The utilities for states 1 and 2 are specified as in Figure 2(c) with the exception that each utility is now a function of a single variable denoted by $x_C$, $x_W$, and $z$. State 1's latent variable equation is thus,

$$y_A^* = p_{\neg R}\beta_C x_C + p_R \beta_W x_W + \epsilon_1,$$

and state 2's latent variable equation is

$$y_R^* = \gamma z + \epsilon_2.$$

For each monte carlo replication, $x_C$, $x_W$, and $z$ are drawn anew from uniform distributions with means of -0.5 and variances of one.[13] The stochastic components $\epsilon_1$ and $\epsilon_2$ are drawn anew from independent normal distributions with means of zero and variances of $\sigma_\epsilon^2$. All coefficients, $\beta_C$, $\beta_W$, and $\gamma$, are set to 1.

The values taken by the two latent variables, $y_A^*$ and $y_R^*$, determine the actions

[13]Using uniform distributions with slightly negative means ensures that war is a relatively rare event in the simulated data. In this way, the data more closely approximate what we find in real-world applications.

taken by both states in the simulated data. State 1 attacks when $y_A^* > 0$, and state 2 resists when $y_R^* > 0$. For each replication, two versions of the dependent variable are generated: one contains the actions of both states and the other is aggregated to war and no war as noted in Section 2.3.1. In this way, we can discriminate between the strategic and selection models and between the strategic and probit models using the same simulated independent variables and error terms.

The strategic, selection, and probit models are estimated for each generated data set. The specification of the strategic model matches the DGP. The selection model is specified with the following selection and outcome equations,

$$\begin{aligned} y_A^* &= \beta_C x_C + \beta_W x_W + \gamma z + \epsilon_1 \\ y_R^* &= \gamma z + \epsilon_2. \end{aligned}$$

The $z$ regressor is added to the selection equation in order that the model might stand a better chance of approximating the strategic DGP.[14]

The probit model is specified as,

$$y_{War}^* = \beta_C x_C + \beta_W x_W + \gamma z + \epsilon.$$

To compare the strategic model to the probit model, log-likelihoods for the

[14]It seems likely that some researchers might try to model state 1's conditioning on state 2's behavior by including $z$ in state 1's regression equation.

strategic model are constructed using the estimated model's parameters and aggregating the probabilities appropriately as in Equations 8 and 9.

Our ability to discriminate between these rival models is likely to depend upon the size of the sample and the "signal-to-noise ratio" of the DGP (the ratio of the variance of the systematic portion of the DGP to the variance of the error term). Discrimination should be easier as both the size of the sample and as the "signal-to-noise ratio" increase. To assess these effects, we varied the size of the sample between 50 and 500, and varied the "signal-to-noise ratio" by changing the error variance between 0.5 and 2.0. 8000 replications are performed.

The following summarizes the experiments:

- Data generating process: strategic (all coefficients set to 1),
- Sample sizes: $N \in \{50, 100, 200, 300, 500\}$,
- Error variance: $\sigma_\epsilon \in \{.5, 1, 2\}$,
- Comparisons: probit v. strategic, selection v. strategic[15],
- Tests: Vuong and distribution-free,
- Replications: 8000.

In total, the results of 15 simulations are reported.

The experimental design raises two interesting issues. First, given the design of the simulations, we cannot discuss the size of the tests because the null

[15] As both the probit and selection models are misspecified given the DGP, how well they fare against each other is not a concern of this paper.

hypothesis is false in every experiment (the models are never equally close to the true DGP). Rejecting the null hypothesis when it is true is therefore not possible. We can, however, discuss the power of the tests in both the correct direction (toward the strategic model) and the wrong direction (away from the strategic model). Note that this latter category includes not only picking the wrong model, but also picking *neither* model.

Second, we are comparing a continuous test statistic, the Vuong, with a discrete test statistic, the number of positive differences. The problem with this comparison is that for any finite number of observations, the exact significance level of the discrete test statistic is unlikely to match the nominal significance level selected for the simulation. For example, we would like to assess the statistics based on a .05 significance level. However, the discrete test statistic has a limited number of probabilities (the number of "jump points" in the c.d.f.) that can serve as $\alpha$. Absent identical exact significance levels, power comparisons may be quite misleading (Gibbons and Chakraborti 1992).

One way to avoid this problem is to employ a randomized decision rule (Lehmann 1986). However, as DeGroot (1989) points out, it seems odd for a researcher to decide which hypothesis to accept by tossing a coin or using some other method of randomization. In place of a randomized procedure, then, we chose critical values for the Vuong test such that the significance level of the Vuong would match the exact significance level of the distribution-free test for the desired $\alpha$. For example, with a sample size of 200, there is no critical value for the binomial that will produce a significance level of .05. Using 58 as a critical value produces a significance level of .0666. The appropriate critical value for the Vuong test, then, is one that also produces a significance level of .0666, which in this case, is 1.5015606. The power levels we report, therefore, are for equivalent nominal and exact significance levels.

### 4.2 Results

The results for the discrimination of the strategic model against the selection model are displayed in Table 1, and the results for the discrimination of the binary strategic model against the probit model are shown in Table 2. Each table reports in what proportion of replications the Vuong and distribution-free (Clarke) tests correctly chose the strategic model.[16] These results are shown for sample sizes ranging from 50 to 500 and for disturbance standard deviations of .5, 1, and 2.

Tables 1 and 2 clearly show that both tests are able to discriminate between the models, depending on the sample size and signal-to-noise ratio. In general, the power of both tests increases as the sample size increases and as the signal-to-noise ratio increases. The former is not surprising as both tests

[16]The strategic and probit models almost always converged. Therefore, the results reported for the binary data model comparisons are generally based on a full 8000 iterations, or at least very close to it. Unfortunately, the selection model had difficulty converging at times for smaller sample sizes and smaller error variances. In these cases, we report the results of only those iterations that converged without problem.

Table 1: Discriminating between the Strategic vs. Selection Models.
*The table displays the proportion of times the strategic model was correctly chosen by the Clarke and Vuong tests. The experiments were conducted for sample sizes ranging from N=50 to N=500, and for disturbance standard deviations ranging from .5 to 2.*

| Size | Test | 0.5 | 1.0 | 2.0 |
|---|---|---|---|---|
| 50 | Clarke | 1.000 | 0.919 | 0.728 |
| | Vuong | 0.979 | 0.775 | 0.688 |
| 100 | Clarke | 1.000 | 0.969 | 0.806 |
| | Vuong | 0.996 | 0.856 | 0.788 |
| 200 | Clarke | 1.000 | 0.984 | 0.855 |
| | Vuong | 1.000 | 0.926 | 0.835 |
| 300 | Clarke | 1.000 | 0.994 | 0.882 |
| | Vuong | 1.000 | 0.974 | 0.873 |
| 500 | Clarke | 1.000 | 0.999 | 0.889 |
| | Vuong | 1.000 | 0.995 | 0.885 |

are consistent and will choose the correct model more often for larger sample sizes. The latter is hardly surprising as the discrimination tests perform better in the absence of noise.

The tests perform at their worst when the sample size is small (N=50) and the uncertainty is large ($\sigma_\epsilon = 2$). In this situation, the distribution-free test correctly selects the strategic model 72.8% of the time, while the Vuong test correctly selects the strategic model 68.8% of the time. The results are even less impressive when we turn to the binary model comparison. Here,

Table 2: Discriminating between the Strategic vs. Probit Models.
*The table displays the proportion of times the strategic model was correctly chosen by the Clarke and Vuong tests. The experiments were conducted for sample sizes ranging from N=50 to N=500, and for disturbance standard deviations ranging from .5 to 2.*

| Size | Test | 0.5 | 1.0 | 2.0 |
|---|---|---|---|---|
| 50 | Clarke | 0.860 | 0.713 | 0.631 |
| | Vuong | 0.699 | 0.484 | 0.351 |
| 100 | Clarke | 0.961 | 0.863 | 0.835 |
| | Vuong | 0.916 | 0.765 | 0.626 |
| 200 | Clarke | 0.996 | 0.967 | 0.952 |
| | Vuong | 0.993 | 0.942 | 0.868 |
| 300 | Clarke | 0.999 | 0.993 | 0.992 |
| | Vuong | 0.999 | 0.993 | 0.972 |
| 500 | Clarke | 1.000 | 0.999 | 0.999 |
| | Vuong | 1.000 | 0.999 | 0.998 |

the distribution-free test correctly chooses the strategic model 63.1% of the time, while the Vuong test only selects the correct model in 35.1% of the iterations.

Large sample sizes, however, are not always required for accurate model discrimination. For example, Table 1 shows that the distribution-free test is highly accurate when the uncertainty is low to moderate, even for very small samples.

Clarke (2002) demonstrates that the distribution-free test generally outperforms the Vuong test for small samples, and performs equally well as samples became relatively large. Tables 1 and 2 provide further evidence of this result. In every case, the distribution-free test performs at least as well as the Vuong test. For small samples, it often performs much better. The greater relative power of the distribution-free test does not, however, come without a price. If we disaggregate the "incorrect" categories in both tables into "chose incorrectly" and "made no choice" (not presented here), we see that the distribution-free test has a slightly higher probability of choosing the wrong model, while the Vuong test, on the other hand, has a slightly higher probability of choosing neither model. We believe that the benefits gained from the greater power of the distribution-free test outweigh the slightly higher probability of rejecting the null in favor of the incorrect model.[17]

The simulation results should be of great interest to substantive scholars. The results are important in that small sample studies, though not the majority, are common in international relations research. For example, 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. A test that works under conditions where is discrimination is difficult is surely welcome.

[17]See Clarke (2003b) for a formal comparison of the trade-offs between these errors.

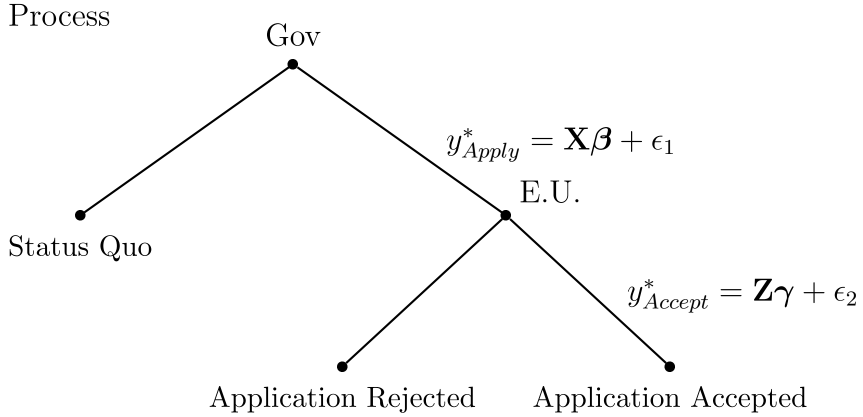
# 5 Application

In their recent article on European Union (EU) eastern enlargement, Plümper, Schneider, and Troeger (2006) break new ground by treating enlargement as two interrelated decisions. In the first stage, the governments of transition states choose whether or not to apply for membership, and in the second stage, the EU decides whether or not to accept these applications. The authors state that the presence of the selection mechanism calls for the use of a Heckman selection model arguing that "the procedure not only captures our theory appropriately, it is also more efficient and robust than competing procedures and — most importantly — the only consistent estimator given the truncated distribution of the sample in the second stage" (Plümper *et al.* 2006, 24).

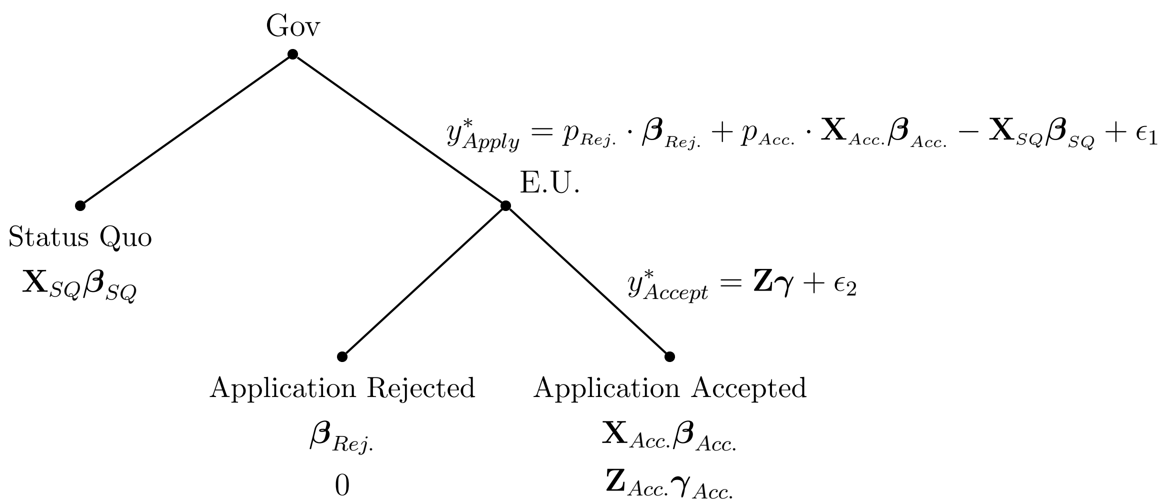
Does the Heckman selection model appropriately capture Plümper *et al.*'s theory? The authors argue that the application and the accession stages are "intertwined" and interdependent [18]. They also argue that leaders of autocratic regimes are unwilling to apply for EU membership because, among other reasons, "they anticipated little or no chance for success" [18]. This suggests that the decision to apply for EU membership is strategic. If so, then, as we argued in Section 2.3, the Heckman selection model does not appropriately capture the theory.

Using the methods discussed in Section 3, we can test whether or not the decision to apply for EU membership is strategic. The two competing models

Figure 3: Modelling EU Enlargement as a Selection Process and a Strategic Process




(a) Selection Model




(b) Strategic Model

are depicted in Figure 3 (compare with panels (b) and (c) of Figure 2). In the top panel of Figure 3, the transitioning government makes the decision to apply for EU membership or not. This decision would be made, as suggested by Plümper *et al.* [21], for purely domestic reasons, and the likely decision by the EU is not a factor. In the bottom panel of Figure 3, the transitioning government makes the decision to apply for EU membership based partly on what the government believes the EU's decision will be. Which model is closer to the DGP, therefore, has considerable theoretical bite.

We replicate Plümper *et al.*'s model 3 (their full model) in Table 3.[18] The dependent variable is binary taking a value of 1 if a state applies for membership in the EU. The independent variables in this stage include a measure of regulatory quality, the Polity IV proxy for the level of democracy, a measure of the dependence of the government on presidential intervention, exports of goods and services (per cent of GDP), and government consumption expenditure (per cent of GDP).

The dependent variable in the second stage is also binary and takes a value of 1 if a state has been selected for accession. In the second stage, the

[18]Our replication is based on 98 observations, not the 169 observations used by Plümper *et al.* The reason for this discrepancy is the amount of missing data in the variables used for the EU utilities. Unlike the selection model, where the two equations are essentially estimated separately, the strategic model uses the variables in the EU utilities to create the probabilities that weight the choice between status quo and applying. The three variables that comprise the EU utilities—chapters completed, the power of eurosceptic parties, and the total size of the agricultural sector—include significant amounts of missing data. Thus, the weighting probabilities have a significant amount of missing data, and we end up with 98 observations. We ran the selection model with 98 observations, therefore, to ensure comparability with the strategic model. A few variables are no longer significant, but no major inferences are affected.

Table 3: Heckman Probit Selection Model, $n = 98$

| Variable | Coefficient | S.E. | $P$-value |
|---|---|---|---|
| *First stage: Gov.* | | | |
| Regulatory quality | 2.262 | 0.444 | 0.000 |
| Level of democracy | -0.281 | 0.271 | 0.301 |
| Presidential power | 0.220 | 0.309 | 0.477 |
| Export of goods and services | -0.009 | 0.013 | 0.474 |
| Government consumption | -0.073 | 0.049 | 0.138 |
| Constant | -2.050 | 2.060 | 0.320 |
| *Second stage: E.U.* | | | |
| Chapter | 0.260 | 0.057 | 0.000 |
| Power of Eurosceptic parties | -0.008 | 0.006 | 0.156 |
| Agriculture | -0.487 | 0.001 | 0.000 |
| $\rho$ | -1.000 | 0.000 | |
| Log-Likelihood | -46.266 | | |

authors "focus on the effect of political variables providing EU members with information on the political preferences of parties in the applicants' parliaments" (Plümper *et al.* 2006, 25). The independent variables therefore include the number of closed chapters in the negotiations and the strength of Eurosceptic parties. Finally, the authors include the total size of the agricultural sector as a proxy for "the expected redistribution of agricultural subsidies once applicants are accepted" [26].

In Table 4, we present a strategic version of EU enlargement. One question that arises is how to apportion to independent variables in the first

Table 4: Strategic Model, $n = 98$

| Variable | Coefficient | S.E. | $P$-value |
|---|---|---|---|
| Government | | | |
| *Status Quo* ($\boldsymbol{\beta}_{SQ}$) | | | |
| Level of Democracy | 1.23 | 0.55 | 0.025 |
| Regulatory Quality | -6.23 | 1.49 | 0.000 |
| *Application rejected* ($\boldsymbol{\beta}_{Rej.}$) | | | |
| Constant | -9.79 | 4.99 | 0.050 |
| *Application accepted* ($\boldsymbol{\beta}_{Acc.}$) | | | |
| Government consumption | 0.03 | 1.28 | 0.982 |
| Presidential power | 44.67 | 32.85 | 0.174 |
| Export of goods and services | -0.71 | 0.64 | 0.269 |
| European Union | | | |
| *Application accepted* ($\boldsymbol{\gamma}_{Acc.}$) | | | |
| Chapter | 0.43 | 0.89 | 0.000 |
| Power of Eurosceptic parties | -0.02 | 0.01 | 0.014 |
| Agriculture | -0.67 | 0.04 | 0.079 |
| Constant | -1.89 | 0.47 | 0.000 |
| Log-Likelihood | -38.49 | | |

stage of the selection model between the three government utilities (status quo, application rejected, application accepted) in the strategic model. Our reasoning is quite straightforward. Reforms must be carried out before the admission decision; thus, they are sunk costs that do not depend on whether the EU admits the state or not. Hence, regulatory quality and regime type should affect the status quo utility of the government. We normalize the payoff to the government for "application rejected" to only a constant. The other independent variables are included in the acceptance utility for the government. For the EU's utilities, we normalize the payoff for "application rejected" to zero and enter the independent variables into the EU's utility for "application accepted."

Comparing the log-likelihoods from Tables 3 and 4, we see that the strategic model appears to fit better, although we do not know whether this difference is meaningful or not. To answer that question, we apply the nonnested tests from Section 3. These results are in Table 5.

Table 5: Results of the Nonnested Tests, $n = 98$

| | Statistic | P-value |
|---|---|---|
| Vuong | 2.01 | 0.045 |
| Clarke | 68 | 0.000 |

The results of the tests leave no room for ambiguity; both tests reject the null hypothesis of equality in favor of the strategic model. Thus, we can be confident that when states decide to apply for EU membership, they pay attention not solely to domestic political concerns, but also to the anticipated decision of the EU itself.

# 6 Conclusion

The purpose of this paper is to demonstrate that discrimination between discrete choice models with different functional forms is possible, even with small samples. We provide a framework in which it is possible to compare strategic models to nonstrategic alternatives, or even strategic models against one another. At the same time, we extend nonnested model testing in political science to situations where the rival models are nonnested in terms of their functional forms.

We demonstrate that discriminating between strategic choice models and various alternative nonstrategic choice models is feasible even under adverse conditions. While the distribution-free test has greater relative power in many of the experiments, both tests perform well and are easy to implement. There is therefore no reason why a substantively-oriented scholar should need to simply assume whether or not that functional form is strategic. In our application, we demonstrate that the two-stage process of EU enlargement is indeed strategic. We hope that future scholars will use these results and techniques for increasingly rigorous comparative model testing.

# References


Akaike, H. 1973. "Information Theory and an Extension of the Likelihood Ratio Principle." In *Second International Symposium of Information Theory*, eds. B.N. Petrov and F. Csaki. Minnesota Studies in the Philosophy of Science, Budapest: Akademinai Kiado.

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. 2001. "Testing Nonnested Models of International Relations: Reevaluating Realism." *American Journal of Political Science* 45:724–744.

Clarke, Kevin A. 2003a. "Nonparametric Model Discrimination in International Relations." *Journal of Conflict Resolution* 47:72–93.

Clarke, Kevin A. 2003b. "A Simple Distribution-Free Test for Nonnested Hypotheses."

DeGroot, Morris H. 1989. *Probability and Statistics*. 2 ed. Reading, MA: Addison-Wesley.

Gibbons, Jean Dickinson, and Subhabrata Chakraborti. 1992. *Nonparametric Statistical Inference*. 3 ed. New York: Marcel Dekker, Inc.

Greene, William H. 2003. *Econometric Analysis*. 5 ed. New Jersey: Prentice Hall.

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 K. 1988. *Extended Deterrence and the Prevention of War*. New Haven: Yale University Press.

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

Kullback, Solomon, and R.A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22:79–86.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. 2 ed. New York: John Wiley.

McKelvey, Richard D., and Thomas R. Palfrey. 1998. "Quantal Response Equilibria for Extensive Form Games." *Experimental Economics* 1:9–41.

Morton, Rebecca B. 1999. *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science*. Cambridge: Cambridge University Press.

Plümper, Thomas, Christina J. Schneider, and Vera E. Troeger. 2006. "The Politics of00 EU Eastern Enlargement: Evidence from a Heckman Selection Model." *British Journal of Political Science* 36:17–38.

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.

Reed, William. 2000. "A Unified Statistical Model of Conflict Onset and Escalation." *American Journal of Political Science* 44:84–93.

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

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

Signorino, Curtis S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93:411–433.

Signorino, Curtis S. 2002. "Strategy and Selection in International Relations." *International Interactions* 28:93–115.

Signorino, Curtis S., and Ahmer Tarar. 2002. "A Unified Theory and Test of Extended Immediate Deterrence."

Signorino, Curtis S., and Kuzey Yilmaz. 2003. "Strategic Misspecification in Regression Models." *American Journal of Political Science* 47:551–566.

Vuong, Quang. 1989. "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica* 57:307–333.