Political Studies

Political Studies Association
60 years of political studies

POLITICAL STUDIES: 2010 VOL 58, 368–388

doi: 10.1111/j.1467-9248.2009.00813.x

Discriminating Methods: Tests for Non-nested Discrete Choice Models

Kevin A. Clarke and Curtis S. Signorino

University of Rochester

We consider the problem of choosing between rival statistical models that are non-nested 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 non-strategic discrete choice models. The distribution-free test appears to have greater relative power in small samples.

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 modeling 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 non-strategic models. While it is clear that strategic specifications provide different answers from traditional specifications, it is 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 non-nested (neither model is a special case of the other model).[1] Discriminating between non-nested models requires specialized tests that are rarely used in political science research. Clarke (2001) introduced the issue of non-nested testing to political science, and Clarke

(2003) introduced a simple distribution-free test for non-nested model discrimination. These articles, however, consider only models that are non-nested in terms of their covariates. Testing models that are non-nested in terms of their functional forms is a natural extension of this line of research.

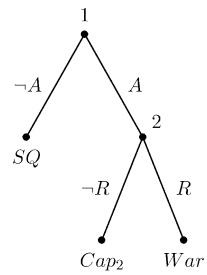
In this article, we demonstrate that discriminating between discrete choice models with different functional forms is possible, even with small samples. In the next section, we present a common crisis scenario and consider three functional forms a researcher might choose when modeling it: a probit model, 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.

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 article. In this crisis situation, state 1 must decide whether to attack (A) state 2 or not attack ($\neg A$). If attacked, state 2 must then choose 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 state 1 attacks and state 2 resists, we assume that war (War) is 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

Figure 1: Conflict Model



State 1 decides whether to attack (A) state 2 or not attack ($\neg A$). If attacked, state 2 decides whether to resist (R) or not resist ($\neg R$). The states' actions lead to three outcomes: the Status Quo (SQ), Capitulation by state 2 (Cap_2) or War (War)

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 from 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 researcher might choose when empirically analyzing the conflict scenario in Figure 1. We choose to 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.

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 modeling choice, two of the outcomes in the conflict scenario, status quo (*SQ*) and capitulation by state 2 (Cap_2), are aggregated into a single outcome, the absence of war ($\neg War$).

Figure 2a 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, $y^{\star}_{War}$, is a linear function of a set of regressors pertaining to state 1 and of a set of regressors pertaining to state 2,

$$y^{\star}_{War} = X\beta + Z\gamma + \varepsilon.$$

X in the above equation represents state 1's regressors, Z represents state 2's regressors, β and γ are coefficient vectors on X and Z, respectively, and ε is a random disturbance, assumed to be normally distributed with mean zero and variance one.[4]

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

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

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

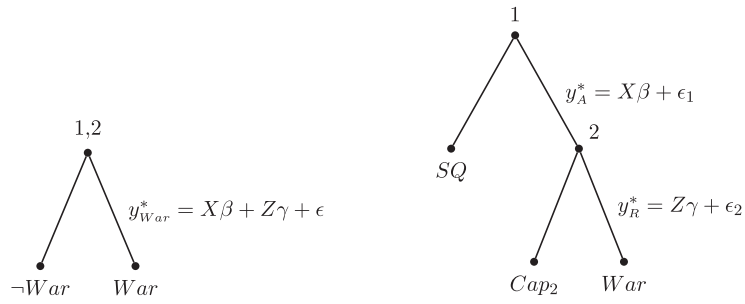
$$\Pr(y_{War} = 0 | X, Z) = 1 - \Phi(X\beta + Z\gamma) \quad (1)$$

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

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

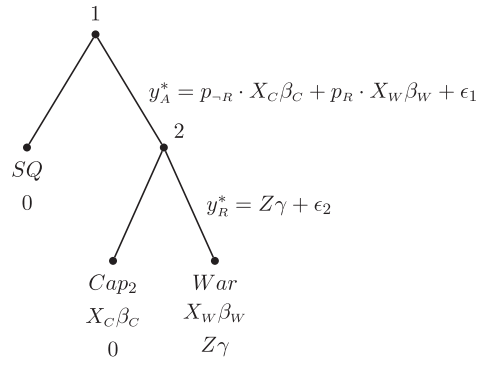
Throughout the remainder of the article, we motivate the statistical models by making random utility assumptions, as opposed to relying on the structure of the data. It is important to note that probit models may also be motivated by random utility assumptions

Figure 2: Alternative Discrete Choice Specifications



(a) Probit Model

(b) Selection Model



(c) Strategic Model

Notes: (a) shows the common probit specification. In the selection model, (b), the War outcome results from a 'selection' equation, y^*_A, and an 'outcome' equation, y^*_R, with the additional assumption that ε_1 and ε_2 are correlated. Finally, (c) displays a strategic model, with each player's pay-offs 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.

(Judge *et al.*, 1985). In random utility models, decision makers are assumed to have preferences over outcomes, which are represented by their utilities for those outcomes. 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]

That said, 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 ¬*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* vs. ¬*War* outcomes. With that said, we turn to two models that can be more directly motivated by random utility assumptions.

Selection Model

Suppose now that a researcher has sequential data on state 1's decision to attack and state 2's decision to resist after being attacked. One random utility-based modeling 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 2b 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',

$$y_A^\star = X\beta + \varepsilon_1 \tag{3}$$

represents state 1's net utility for attacking, where $X\beta$ is the observable component of its utility and ε_1 is the random term.[7] As a utility maximizer, state 1 attacks when $y_A^\star > 0$. We observe,

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

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 vs. capitulating. The 'outcome equation',

$$y_R^\star = Z\gamma + \varepsilon_2,$$

represents state 2's net utility for resisting, where $Z\gamma$ is the observable component of its utility and ε_2 is the random term. As a utilty maximizer, state 2 resists whenever $y_R^\star > 0$. We observe,

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

The final step in specifying the selection model concerns the disturbances, ε_1 and ε_2. Following William Greene (2003), we assume the disturbances are distributed bivariate normal with mean zero, variance one and correlation ρ. Given these assumptions, the probability of observing each outcome is,

$$\Pr(SQ) = \Pr(\gamma_A = 0|X, Z) = 1 - \Phi(X\beta)$$

$$\Pr(Cap_2) = \Pr(\gamma_A = 1, \gamma_R = 0|X, Z) = \Phi_2(X\beta, -Z\gamma, -\rho)$$

$$\Pr(War) = \Pr(\gamma_A = 1, \gamma_R = 1|X, Z) = \Phi_2(X\beta, Z\gamma, \rho),$$

where $\Phi(\cdot)$ and $\Phi_2(\cdot)$ are the CDFs of the standard normal and standard bivariate normal, respectively, and ρ is the correlation between the error terms for the two equations.

Strategic Choice Model

A third modeling 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 2c 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 pay-offs to each state are given below the outcomes in Figure 2c. We normalize the status quo pay-off 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 pay-off for the capitulation outcome ($X_C\beta_C$) and (2) those that affect state 1's pay-off for the war outcome ($X_W\beta_W$). As before, we normalize state 2's pay-off for the status quo at zero, and we let its pay-off 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,

$$\gamma_R^\star = Z\gamma + \varepsilon_2,$$

represents state 2's net utility for resisting. $Z\gamma$ is the observable component of the utility, and ε_2 is the random term. As a utility maximizer, state 2 resists whenever $\gamma_R^\star > 0$. We observe,

$$\gamma_R = \begin{cases} 1, & \textit{if } \gamma_R^\star > 0 \\ 0, & \textit{otherwise}. \end{cases}$$

Given the distributional assumption for ε_2, state 2's choice probabilities are,

$$p_R = \Pr(\gamma_R = 1|X, Z) = \Phi(Z\gamma)$$

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

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 vs. 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. 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,

$$y_A^\star = p_{\neg R} \cdot X_C\beta_C + p_R \cdot X_W\beta_W + \varepsilon_1, \tag{4}$$

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 ε_1 is the random utility component. State 1 attacks when $y_A^\star > 0$. We observe,

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

State 1's equilibrium choice probabilities are then,

$$p_A = \Pr(y_A = 1|X, Z) = \Phi(p_{\neg R} \cdot X_C\beta_C + p_R \cdot X_W\beta_W)$$

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

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

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

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

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

Maximum likelihood estimation of the effect parameters, β_C, β_W and γ, is based on these equilibrium probabilities assuming that the dependent variable denotes which of the three outcomes occurred for each observation.

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 2a that the binary data represent *War* vs. $\neg War$. 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.

Non-nested Model Testing

The models in the previous section are non-nested 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 controversial of these tests are the Vuong test (Vuong, 1989) and a distribution-free test introduced by Clarke (2003).

Both tests are based on the Kullback-Leibler information criteria (KLIC) (Kullback and Leibler, 1951). Quang Vuong (1989) defines the KLIC as,

$$KLIC \equiv E_0[\ln h_0(Y_i|X_i)] - E_0[\ln f(Y_i|X_i;\beta_\star)],$$

where $h_0(.|.)$ is the true conditional density of Y_i given X_i (that is, the true but unknown model), E_0 is the expectation under the true model and $\beta_\star$ are the pseudo-true values of β (the estimates of β when $f(Y_i|X_i)$ is not the true model). The best model is the model that minimizes the KLIC, for the best model is the one that is closest to the true specification. We should therefore choose the model that maximizes $E_0[\ln f(Y_i|X_i;\beta_\star)]$. In other words, one model should be selected over another if the individual log-likelihoods of that model are significantly larger than the individual log-likelihoods of the rival model.

The Vuong Test

The null hypothesis of Vuong's test is,

$$H_0 : E_0\left[\ln\frac{f(Y_i|X_i;\beta_\star)}{g(Y_i|Z_i;\gamma_\star)}\right] = 0,$$

which states that the two models are equally close to the true specification.[10]

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

$$\frac{1}{n} LR_n\left(\hat{\beta}_n, \hat{\gamma}_n\right) \xrightarrow{a.s.} E_0\left[\ln \frac{f\left(Y_i \middle| X_i; \beta_\star\right)}{g\left(Y_i \middle| Z_i; \gamma_\star\right)}\right],$$

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

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

where

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

and

$$\hat{\omega}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\ln \frac{f\left(Y_i \middle| X_i; \hat{\beta}_n\right)}{g\left(Y_i \middle| Z_i; \hat{\gamma}_n\right)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} \ln \frac{f\left(Y_i \middle| X_i; \hat{\beta}_n\right)}{g\left(Y_i \middle| Z_i; \hat{\gamma}_n\right)}\right]^2.$$

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

The log-likelihoods used in 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 Hirotsugu Akaike's (1973) information criteria or Gideon Schwarz's (1978) Bayesian information criteria. In the simulations that follow, we use the latter, making the adjusted statistic,[11]

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

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

The Distribution-Free Test

The Vuong test is not an exact test; it is normally distributed asymptotically. Simulations demonstrate that the relatively small sample sizes used in some international relations

research present a problem for the power of the test (Clarke, 2003). A brief explanation provides some intuition.[12] As stated above, under the null hypothesis, the Vuong statistic is distributed as a standard normal,

$$\frac{LR_n\left(\hat{\beta}_n, \hat{\gamma}_n\right)}{\left(\sqrt{n}\right)\hat{\omega}_n} \xrightarrow{D} N(0,1).$$

An equivalent asymptotic result (Greene, 2003) is that the mean log-likelihood ratio converges almost surely to a normal distribution with mean 0 and asymptotic variance $\hat{\omega}_n^2/n$,

$$\frac{1}{n} LR_n\left(\hat{\beta}_n, \hat{\gamma}_n\right) \xrightarrow{a.s.} N\left(0, \frac{\hat{\omega}_n^2}{n}\right).$$

This convergence, however, is quite slow. For sample sizes under 500, the distribution is highly leptokurtic – very much like a double-exponential distribution. As Erich Lehmann (1986) points out, the sign test is the LMP test for testing $\theta \leq 0$ against $\theta > 0$ when the sample is drawn from a double-exponential distribution. The sign test therefore seems to be the obvious solution for situations in which only small-to-modest sample sizes are available.

Clarke's (2007) distribution-free test applies a modified paired sign test to the differences in the individual log-likelihoods from two non-nested models. While the Vuong test determines whether or not the average log-likelihood ratio is statistically different from zero, the proposed test determines whether or not the *median* log-likelihood ratio is statistically different from zero. If the models are equally close to the true specification, half 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 the individual log-likelihood ratios should be greater than zero. Conversely, if model *g* is 'better' than model *f*, more than half the individual log-likelihood ratios should be less than zero.

Utilizing Vuong's notation, the null hypothesis of the distribution-free test is:

$$H_0 : \Pr\left(\ln f\left(Y_i \middle| X_i; \beta_\star\right) > \ln g\left(Y_i \middle| Z_i; \gamma_\star\right)\right) = 0.5$$

$$\Pr\left(\ln f\left(Y_i \middle| X_i; \beta_\star\right) < \ln g\left(Y_i \middle| Z_i; \gamma_\star\right)\right) = 0.5$$

or, equivalently,

$$H_0 : \Pr\left(\ln f\left(Y_i \middle| X_i; \beta_\star\right) - \ln g\left(Y_i \middle| Z_i; \gamma_\star\right) > 0\right) = 0.5$$

$$\Pr\left(\ln f\left(Y_i \middle| X_i; \beta_\star\right) - \ln g\left(Y_i \middle| Z_i; \gamma_\star\right) < 0\right) = 0.5$$

The assumptions of the test are unsurprising and quite general. First, the differences, $\ln \frac{f\left(Y_i \middle| X_i; \beta_\star\right)}{g\left(Y_i \middle| Z_i; \gamma_\star\right)}$, are mutually independent.[13] Second, each $\ln \frac{f\left(Y_i \middle| X_i; \beta_\star\right)}{g\left(Y_i \middle| Z_i; \gamma_\star\right)}$ comes from a continuous population (not necessarily the same) that has a common median θ.[14] The

consistency of the test is established simply by strengthening the second assumption to require that each difference has the same continuous population with median θ (Hollander and Wolfe, 1999). Proofs of consistency and unbiasedness for the distribution-free test are in the Appendix.

Letting $Z_i = \ln f(Y_i|X_i;\hat{\beta}_n) - \ln g(Y_i|Z_i;\hat{\gamma}_n)$, and

$$\Psi_i = \begin{cases} 1 & if\ Z_i > 0 \\ 0 & if\ Z_i < 0, \end{cases}$$

the test statistic is

$$B = \sum_{i=1}^{n} \Psi_i.^{15}$$

One of the great strengths of this procedure is that implementation is remarkably simple; the test can be produced by any mainstream statistical software package using the following algorithm:[16]

(1) Run model *f*, saving the individual log-likelihoods, $\ln f(Y_i|X_i;\hat{\beta}_n)$;
(2) run model *g*, saving the individual log-likelihoods, $\ln g(Y_i|Z_i;\hat{\gamma}_n)$;
(3) compute the differences, Z_i, and count the number of positive values, Ψ_i;
(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,

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

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:

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

While we cannot justify any particular correction, we can broadly justify the approach by appealing to Vuong's justification for his correction. Vuong notes that as long as the correction factor divided by the square root of *n* has a stochastic order of 1,

$$n^{-1/2}K_n(F_\theta, G_\gamma) = o_p(1),$$

the adjusted statistic has the same asymptotic properties of the unadjusted statistic.

Vuong's justification amounts to pointing out that the asymptotic properties of the adjusted statistic are the same as the asymptotic properties of the unadjusted statistic. If we consider the normal approximation to the distribution-free test detailed above,[17] we can see that the asymptotic properties of the test are unaffected by the correction.

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 performed 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.

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 2c 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,

$$y_A^\star = p_{\neg R}\beta_C x_C + p_R\beta_W x_W + \varepsilon_1,$$

and state 2's latent variable equation is

$$y_R^\star = \gamma z + \varepsilon_2.$$

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.[18] The stochastic components ε_1 and ε_2 are drawn anew from independent normal distributions with means of zero and variances of σ_ε^2. All coefficients, β_C, β_W and γ, are set to 1.

The values taken by the two latent variables, $y_A^\star$ and $y_R^\star$, determine the actions taken by both states in the simulated data. State 1 attacks when $y_A^\star > 0$, and state 2 resists when $y_R^\star > 0$. 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 the second section. 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,

$$y_A^\star = \beta_C x_C + \beta_W x_W + \gamma' z + \varepsilon_1$$

$$y_R^\star = \gamma z + \varepsilon_2.$$

The z regressor is added to the selection equation in order that the model might stand a better chance of approximating the strategic DGP.[19]

The probit model is specified as,

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

To compare the strategic model to the probit model, log-likelihoods for the 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 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. Eight thousand replications were 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: $\sigma_\varepsilon \in \{0.5,1,2\}$
- Comparisons: probit vs. strategic, selection vs. strategic[20]
- Tests: Vuong and distribution-free
- Replications: 8,000

In total, the results of fifteen 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 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 0.05 significance level. However, the discrete test statistic has a limited number of probabilities (the number of 'jump points' in the CDF) that

can serve as α. 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 Morris 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 α. For example, with a sample size of 200, there is no critical value for the binomial that will produce a significance level of 0.05. Using 58 as a critical value produces a significance level of 0.0666. The appropriate critical value for the Vuong test, then, is one that also produces a significance level of 0.0666, which in this case is 1.5015606. The power levels we report, therefore, are for equivalent nominal and exact significance levels.

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.[21] These results are shown for sample sizes ranging from 50 to 500 and for disturbance standard deviations of 0.5, 1 and 2.

Table 1: Discrimination between the Strategic vs. Selection Models

		Standard deviation of the error		
Size	*Test*	*0.5*	*1*	*2*
50	Clarke	1	0.919	0.728
	Vuong	0.979	0.775	0.688
100	Clarke	1	0.969	0.806
	Vuong	0.996	0.856	0.788
200	Clarke	1	0.984	0.855
	Vuong	1	0.926	0.835
300	Clarke	1	0.994	0.882
	Vuong	1	0.974	0.873
500	Clarke	1	0.999	0.889
	Vuong	1	0.995	0.885

Notes: 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 0.5 to 2.

Table 2: Discrimination between the Strategic vs. Probit Models

		Standard Deviation of the error		
Size	*Test*	*0.5*	*1*	*2*
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	0.999	0.999
	Vuong	1	0.999	0.998

Notes: 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 0.5 to 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 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 ($\sigma_\varepsilon = 2$). In this situation, the distribution-free test correctly selects the strategic model 72.8 per cent of the time, while the Vuong test correctly selects the strategic model 68.8 per cent of the time. The results are even less impressive when we turn to the binary model comparison. Here, the distribution-free test correctly chooses the strategic model 63.1 per cent of the time, while the Vuong test only selects the correct model in 35.1 per cent 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 (2003) demonstrates that the distribution-free test generally outperforms the Vuong test for small samples, and performs equally well as samples become 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.[22]

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 *et al.* (1993), which has an *n* of 97; Reiter and Stam (1998), which has an *n* of 197; Signorino and Tarar (2006), 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 discrimination is difficult is surely welcome.

Conclusion

The purpose of this article 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 non-strategic alternatives, or even strategic models against one another. At the same time, we extend non-nested model testing in political science to situations where the rival models are non-nested in terms of their functional forms.

We demonstrate that discriminating between strategic choice models and various alternative non-strategic 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 simply to assume whether or not that functional form is strategic. We hope that future scholars will use these results and techniques for increasingly rigorous comparative model testing.

Appendix: Properties of the Non-parametric Test

Two very intuitive and desirable properties that any useful hypothesis test should possess are consistency and unbiasedness. In the following two sections, these properties are proved for the non-parametric test. The assumptions noted in the third section of the article are assumed to hold.

Consistency

A consistent test is one that rejects a false null hypothesis with probability one asymptotically.

Definition A.1 (Fraser, 1957). If $T_{m,n}$ denotes a sequence of size-α tests of $H_0 : \theta \in \omega$ vs. $H_1 : \theta \in \Omega - \omega$, then the sequence is said to be consistent for $\zeta \subset \Omega -$ *omega* if

$$\lim_{n\to\infty} p_{T_{m,n}}(\theta) = 1$$

for $\theta \in \zeta$.

To prove consistency, we can make use of the following theorem.

Theorem A.1 (Lehmann, 1951). Let $\theta = f(F,G)$ be a real valued function such that $f(F,F) = \theta_0$ for all (F,F) in a class C_0. Let $T_{m,n} = t_{m,n}(X_1,...,X_m,Y_1,...,Y_n)$ be a sequence of real valued statistics such that $T_{m,n}$ tends to θ in probability as $min(m,n) \rightarrow \infty$. Suppose that $f(F,G) > \theta_0 (\neq \theta_0)$ for all (F,G) in a class C_1. Then the sequence of tests that reject when $T_{m,n} - \theta_0 > C_{m,n}$ (when $|T_{m,n} - \theta_0| > C'_{m,n}$) is consistent for testing $H: C_0$ at every fixed level of significance against the alternatives C_1.

To review, the null hypothesis of the test is that,

$$H_0 : \theta = \theta_0.$$

Letting $Z_i = \ln f(Y_i|X_i;\hat{\beta}_n) - \ln g(Y_i|Z_i;\hat{\gamma}_n)$, and

$$\Psi_i = \begin{cases} 1 & if\ Z_i > 0 \\ 0 & if\ Z_i < 0 \end{cases}$$

the test statistic is

$$B = \sum_{i=1}^{n} \Psi_i.$$

In the two-tailed case, we reject when

$$\left| \frac{B}{N} - \theta_0 \right| > C'_{\frac{\alpha}{2}},$$

where $C'_{\frac{\alpha}{2}}$ is the smallest integer satisfying

$$\sum_{C'=C'_{\frac{\alpha}{2}}}^{N} \binom{N}{C'} \theta_0^N \leq \frac{\alpha}{2}$$

Applying Theorem A.1, the expected value under the null is,

$$\begin{aligned} E_0\left[\frac{B}{N}\right] &= \frac{1}{N} E_0\left[\sum_{i=1}^{N} \Psi_i\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} E_0[\Psi_i] \\ &= \frac{N\theta_0}{N} \\ &= \theta_0. \end{aligned}$$

The variance is,

$$\begin{aligned} V\left[\frac{B}{N}\right] &= \frac{1}{N^2} V\left[\sum_{i=1}^{N} \Psi_i\right] \\ &= \frac{1}{N^2} \sum_{i=1}^{N} V[\Psi_i] \\ &= \frac{1}{N^2} \sum_{i=1}^{N} E_0[\Psi_i^2] - E_0[\Psi_i]^2 \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \theta_0 - \theta_0^2 \\ &= \frac{N\theta_0(1-\theta_0)}{N^2} \\ &= \frac{\theta_0(1-\theta_0)}{N} \end{aligned}$$

which goes to 0 as $n \to \infty$. We can therefore conclude that B is a consistent test statistic.

Unbiasedness

An unbiased test is one where the probability of rejection under the null hypothesis is never larger than the probability of rejection under the alternative.

Definition A.2 (Fraser, 1957). A test $T_{m,n}$ of $H_0 : \theta \in \omega$ vs. $H_1 \theta \in \Omega - \omega$ is unbiased of size-α if

$$P_{T_{m,n}}(\theta) \geq \alpha$$

for all $\theta \in \Omega - \omega$.

We prove unbiasedness by noting that the non-parametric test reaches its exact significance level for all distributions and that its power function is monotonic. We prove the latter point using the following theorem.

Theorem A.2 (Randles and Wolfe, 1979). Suppose that for testing H_0 vs. H_1 we reject H_0 for large (small) values of a test statistic $T(X_1,\ldots,X_n)$ that satisfies

$$T(x_1 + k, \ldots, x_n + k) \geq (\leq) T(x_1, \ldots, x_n)$$

for every $K \geq 0$ and $(x_1,\ldots,x_n)$. Then the test has a monotone power function in θ for the one-sample location problem; that is,

$$P_T(\theta, F) \leq P_T(\theta', F) \text{ for } \theta \leq \theta',$$

and any continuous distribution with CDF $F(\cdot)$.

The non-parametric test rejects for large (small) values of

$$B(Z_i, \ldots, Z_n) = \left[\sum_{i=1}^{n} \Psi_i > \theta_0\right]$$

where Z_i and Ψ_i are defined as in the first section of the Appendix. When $k > 0$,

$$\begin{aligned} B(z_1 + k, \ldots, z_n + k) &= \left[\sum_{i=1}^{n} (\Psi_i + k) > \theta_0\right] \\ &= \left[\sum_{i=1}^{n} \Psi_i > (\theta_0 - k)\right] \\ &> B(z_1, \ldots, z_n) \end{aligned}$$

The non-parametric is therefore an unbiased test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > (<)\ \theta_0$.

(*Accepted*: 7 August 2009)

About the Authors

Kevin A. Clarke is Associate Professor of Political Science at the University of Rochester. He is a political methodologist with interests in quantitative theory comparison, philosophy of science and international relations. His work has appeared in *American Political Science Review*, *American Journal of Political Science* and *Political Analysis*, among others.

Kevin A. Clarke, Department of Political Science, Harkness Hall, University of Rochester, Rochester, NY 14627-0146, USA; email: *kevin.clarke@rochester.edu*

Curtis S. Signorino is Associate Professor of Political Science at the University of Rochester. In his research he develops statistical methods for analysing strategic decision making and applies those methods to issues in international conflict. His publications have appeared in the *American Political Science Review*, the *American Journal of Political Science*, *International Interactions*, *International Studies Quarterly*, the *Journal of Conflict Resolution* and *Political Analysis*.

Curtis S. Signorino, Department of Political Science, Harkness Hall, University of Rochester, Rochester, NY 14627-0146, USA; email: *curt.signorino@rochester.edu*

Notes

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, and also at the 2008 *Political Studies* Workshop on 'Dialogue and Innovation in Contemporary Political Science'; we thank the participants for their comments. Support from the National Science Foundation (Grants SES-0213771 and SES-0413381) is gratefully acknowledged.

1 See Clarke (2001) for a technical definition of 'non-nested'.

2 The scenario could just as well represent an extended deterrence situation where state 1 is threatening to attack a protégé of state 2. If the protégé is attacked, state 2 must decide whether to defend its protégé (see Huth, 1988; Signorino and Tarar, 2006).

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

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

5 For examples of strategic and non-strategic random utility models, see Signorino (2003).

6 See Greene, 2003; Reed, 2000; Signorino, 2002.

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

8 This is the probit agent error model discussed in more detail in Signorino (2003). For a logit version, see McKelvey and Palfrey (1998); Signorino (1999).

9 See Clarke (2001) for a definition of non-nested and methods of determining whether rival models are non-nested.

10 γ_* and z_i in model g are analogous to β_* and X_i in model f.

11 Which correction factor is used makes no difference to this analysis.

12 See Clarke (2007) for an in-depth explanation.

13 This assumption does not mean that the individual likelihoods themselves must be independent, only that their differences be mutually independent.

14 Hollander and Wolfe (1999) note that when testing $\theta = 0$, this assumption can be weakened to

$$\Pr\left[\ln\frac{f(Y_i|X_i;\beta_\star)}{g(Y_i|Z_i;\gamma_\star)} < 0\right] = \Pr\left[\ln\frac{f(Y_i|X_i;\beta_\star)}{g(Y_i|Z_i;\gamma_\star)} > 0\right] = 0.5.$$

15 To test whether the rival models are different by some non-zero constant, C, simply subtract C from each of the differences and then compute the test statistic (Bradley, 1968).

16 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.

17 The same assumption that provides consistency guarantees asymptotic normality.

18 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.

19 It also seemed likely to us 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.

20 As both the probit and selection models are mis-specified given the DGP, how well they fare against each other is not a concern of this article.

21 The strategic and probit models almost always converged. Therefore, the results reported for the binary data model comparisons are generally based on a full 8,000 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.

22 See Clarke (2007) for a formal comparison of the trade-offs between these errors.

References

Akaike, H. (1973) 'Information Theory and an Extension of the Likelihood Ratio Principle', in B. N. Petrov and F. Csaki (eds), *Second International Symposium of Information Theory*. Minnesota Studies in the Philosophy of Science. Budapest: Akademinai Kiado, pp. 267–81.

Bennett, D. S. and Stam, A. C. (1996) 'The Duration Of Interstate Wars, 1816–1985', *American Political Science Review*, 90 (2), 239–57.

Benoit, K. (1996) 'Democracies Really Are More Pacific (in General): Re-examining Regime Type and War Involvement', *Journal of Conflict Resolution*, 40 (4), 636–57.

Bradley, J. V. (1968) *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ: Prentice-Hall.

Clarke, K. A. (2001) 'Testing Nonnested Models of International Relations: Re-evaluating Realism', *American Journal of Political Science*, 45 (3), 724–44.

Clarke, K. A. (2003) 'Nonparametric Model Discrimination in International Relations', *Journal of Conflict Resolution*, 47 (1), 72–93.

Clarke, K. A. (2007) 'A Simple Distribution-Free Test for Nonnested Hypotheses', *Political Analysis*, 15 (3), 347–63.

DeGroot, M. H. (1989) *Probability and Statistics*, second edition. Reading MA: Addison-Wesley.

Fraser, D. A. S. (1957) *Nonparametric Methods in Statistics*. New York: John Wiley and Sons.

Gibbons, J. D. and Chakraborti, S. (1992) *Nonparametric Statistical Inference*, third edition. New York: Marcel Dekker, Inc.

Greene, W. H. (2003) *Econometric Analysis*, fifth edition. Upper Saddle River, NJ: Prentice Hall.

Hollander, M. and Wolfe, D. A. (1999) *Nonparametric Statistical Methods*, second edition. New York: John Wiley and Sons.

Huth, P. K. (1988) *Extended Deterrence and the Prevention of War*. New Haven CT: Yale University Press.

Huth, P., Christopher, G. and Bennett, D. S. (1993) 'The Escalation of Great Power Militarized Disputes: Testing Rational Deterrence Theory and Structural Realism', *American Political Science Review*, 87 (3), 609–23.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H. and Lee, T.-C. (1985) *The Theory and Practice of Econometrics*, second edition. New York: John Wiley and Sons.

Kullback, S. and Leibler, R. A. (1951) 'On Information and Sufficiency', *Annals of Mathematical Statistics*, 22 (1), 79–86.

Lehmann, E. L. (1951) 'Consistency and Unbiasedness of Certain Nonparametric Tests', *Annals of Mathematical Statistics*, 22 (2), 165–79.

Lehmann, E. L. (1986) *Testing Statistical Hypotheses*, second edition. New York: John Wiley.

McKelvey, R. D. and Palfrey, T. R. (1998) 'Quantal Response Equilibria for Extensive Form Games', *Experimental Economics*, 1 (1), 9–41.

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

Pollins, B. M. (1996) 'Global Political Order, Economic Change, and Armed Conflict: Coevolving Systems and the Use of Force', *American Political Science Review*, 90 (1), 103–17.

Randles, R. H. and Wolfe, D. A. (1979) *Introduction to the Theory of Nonparametric Statistics*. New York: John Wiley and Sons.

Reed, W. (2000) 'A Unified Statistical Model of Conflict Onset and Escalation', *American Journal of Political Science*, 44 (1), 84–93.

Reiter, D. and Stam, A. (1998) 'Democracy, War Initiation, and Victory', *American Political Science Review*, 92 (2), 377–89.

Schwarz, G. (1978) 'Estimating the Dimension of a Model', *Annals of Statistics*, 6 (2), 461–4.

Signorino, C. S. (1999) 'Strategic Interaction and the Statistical Analysis of International Conflict', *American Political Science Review*, 93 (2), 411–33.

Signorino, C. S. (2002) 'Strategy and Selection in International Relations', *International Interactions*, 28 (1), 93–115.

Signorino, C. S. (2003) 'Structure and Uncertainty in Discrete Choice Models', *Political Analysis*, 11 (4), 316–44.

Signorino, C. S. and Tarar, A. (2006) 'A Unified Theory and Test of Extended Immediate Deterrence', *American Journal of Political Science*, 50 (3), 586–605.

Signorino, C. S. and Yilmaz, K. (2003) 'Strategic Misspecification in Regression Models', *American Journal of Political Science*, 47 (3), 551–66.

Vuong, Q. (1989) 'Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses', *Econometrica*, 57 (2), 307–33.