

A Nonparametric Approach to Testing Multiple Competing Models

Kevin A. Clarke[†]
University of Rochester

September 5, 2011

[†]Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: kevin.clarke@rochester.edu. An earlier version of this paper was presented at the 2007 annual meeting of the Political Methodology Society at Pennsylvania State University; I thank the participants for their comments. I also thank John Jackson, David Primo, Randy Stone, Michael Peress, and Larry Rothenberg for helpful comments and discussion. Finally, I thank Craig Volden, Cliff Carrubba, Paul Huth, and Todd Allee for graciously sharing their data. Jeremy Kedziora provided excellent research assistance. Errors remain my own.

Abstract

It is common for political scientists to compare three or more different statistical models in a single article. Rigorously testing multiple models against one another is difficult, and scholars often rely on *ad hoc* procedures. In this paper, we introduce the first model discrimination test for three or more competing nonnested models. The procedure is straightforward and can be implemented using standard statistical software. The models being compared may be nonnested either in their covariates or in their functional forms. We demonstrate the usefulness of the test by considering multiple competing explanations for the democratic peace.

1 Introduction

A theory tested in isolation of its rivals has not truly been tested. Rather, it has survived a kind of plausibility check; the successful theory can be said to provide a plausible interpretation of the data. Most political scientists, however, are interested in making claims about their theories that go beyond mere plausibility. The only way to make such statements is to be comparative. The statistician Paul Rosenbaum (1999), citing Lakatos (1981), Popper (1965) and Feyerabend (1968), points out that theories must be confronted, not with null or alternative hypotheses, but with full-fledged rival theories. Meaningful theory testing requires at least two theories and the data.

Political scientists are embracing the comparative approach. In the American politics subfield, Lawrence, Maltzman, & Smith (2006) compare four competing theories of legislative behavior. In comparative politics, Volden & Carrubba (2004) consider five competing theories of oversized coalition formation. In international relations, Huth & Allee (2002) discriminate between three competing theories of the link between domestic institutions and foreign policy decisionmaking. In each of these cases, the models to be compared are nonnested (the rival models are not special cases of one another), and testing multiple nonnested models against one another is not straightforward.

While classical tests have been proposed for the pairwise comparison of nonnested models (Cox 1961, Vuong 1989, Clarke 2007), no equivalent technique exists for the comparison of multiple nonnested models. The purpose

of this paper is to introduce a procedure for testing three or more nonnested models against one another. The key to the approach is treating model specifications as treatments applied to a set of observations. Viewing specifications as treatments means that the tools used for analyzing multiple related samples (known as randomized complete block designs) can be applied to the discrimination of multiple models. The result is a set of procedures that are simple, possess good statistical properties, take advantage of existing software, and provide researchers with the ability to determine whether a particular model outperforms its rivals.

Unlike much of the recent work on this important problem, which has been largely Bayesian (see George 2000 for an overview), this paper takes a classical (frequentist) hypothesis testing approach for two significant and related reasons. First, although Bayesianism has made inroads among some political methodologists, few substantive political scientists are explicitly Bayesian and classical procedures remain the norm in the discipline (see Efron 1986 for some reasons why statisticians have not adopted Bayesianism wholesale). Second, although Bayesian model selection is theoretically quite simple—by specifying a prior over the set of models and using Bayes’s theorem, the best model is the one with the highest posterior probability—the practical implementation is often difficult and tricky (Chipman, George, & McCulloch 2001). Thus, despite the advent of relatively user-friendly Bayesian software such as WinBugs and MCMCpack (Martin & Quinn 2006), a classical technique will be more widely accepted, and used, than a Bayesian one.

The plan of the paper is as follows. Section 2 demonstrates why pairwise tests and artificial nesting techniques are inadequate approaches to the problem of discriminating between multiple models. Section 3 presents the rationale for viewing model specifications as treatments and multiple model comparisons as randomized complete block designs (RCBDs). Section 4 discusses a nonparametric test for RCBDs known as the Friedman test and shows how it can be adapted for the purpose of model testing. Section 5 discusses the use of multiple comparisons procedures that follow the use of ANOVA-type tests. In concert with the test, these comparisons allow a researcher to rank order the competing models. Section 6 reanalyzes Huth & Allee (2002) and demonstrates that the new test leads to a substantially different conclusion from the one drawn in the literature. Section 7 concludes.

2 Why Multiple Model Tests

It would seem that multiple model tests are not necessary given the availability of three classical tests for pairwise model comparison (the Cox test, the Vuong test, and Clarke's distribution-free test). It appears simpler to perform as many pairwise tests as necessary. For 3 models, that means running 3 tests (the formula is $\binom{n}{k}$, where $k = 2$). For 4 models, the number of pairwise tests goes to 6, and for 5 models, the number goes to 10. The problem, however, is not the number of tests that must be performed. Rather, the problem is what happens to the probability of falsely rejecting the null

hypothesis, that is, a type I error.

By setting the probability of a type I error at 5%, a researcher indicates that she is willing to mistakenly reject the null hypothesis 5% of the time. The more tests performed, however, the less surprising it is to observe something that should happen only 5% of the time.¹ If the tests are mutually independent and the error rate for a single test is α , the probability of at least one type I error in m tests is $1 - (1 - \alpha)^m$. A comparison of 3 models would require 3 tests, and the probability of at least one type I error would be $1 - (1 - 0.05)^3 \approx 0.14$, given $\alpha = 0.05$.

The tests, of course, are not independent as they are performed on the same data. While the actual error rate cannot be computed under these conditions, it is well-known that the maximum error rate is $m * \alpha$, where m is again the number of individual tests that would have to be performed (Miller 1981, Hochberg & Tamhane 1987). Thus, a comparison of 3 models with $\alpha = 0.05$ would have a maximum error rate of $3 * 0.05 = 0.15$. As in this example, the error rate for independent tests is always smaller than the rate for non-independent tests. Table 1 provides type I error probabilities for the comparison of 3 to 6 models.²

If we were to replicate the Volden & Carrubba (2004) study, where 5 models are compared, the probability of at least one type I error in the study would

¹Tukey 1977 refers to this as the problem of “multiplicity.”

²This type of error rate, one that applies to a group of comparisons, is known as a familywise error rate. See Miller 1981 for a discussion of the concept of families.

Table 1: Type I Error Probabilities

# of Models	# of Tests	Probabilities
3	3	$0.14 \leq \Pr(\text{at least one Type I error}) \leq 0.15$
4	6	$0.26 \leq \Pr(\text{at least one Type I error}) \leq 0.30$
5	10	$0.40 \leq \Pr(\text{at least one Type I error}) \leq 0.50$
6	15	$0.53 \leq \Pr(\text{at least one Type I error}) \leq 0.75$

be between 40% and 50%, which is unacceptably high. Simply decreasing the significance level in a set of pairwise tests cannot solve the problem for the simple reason that we do not know the correct level in any particular case. A correct type I error probability can be obtained only using a multiple model test.

Getting the correct type I error probability is not the only reason for using a multiple model test. An equally compelling reason is that the conventional approach to testing multiple models, artificially nesting the rival models in a single larger model and then using a likelihood ratio test, generally fails. Kmenta (1986, 596) and Greene (2003, 154) make this clear with a simple example, which is generalized below to more than two rival models.³

Consider three competing models:

³Although models may be nonnested either in their functional forms or in their covariates, this paper focuses on the latter problem, known as the variable selection problem, as comparing models with different functional forms remains quite rare in political science. The approach taken here applies equally well to either situation.

$$\text{Model 1: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_1,$$

$$\text{Model 2: } \mathbf{y} = \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\epsilon}_2,$$

$$\text{Model 3: } \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_3.$$

Let $\tilde{\mathbf{X}}$ be the variables that appear in \mathbf{X} , but do not appear in \mathbf{W} or \mathbf{Z} , $\tilde{\mathbf{W}}$ be the variables that appear in \mathbf{W} , but do not appear in \mathbf{X} or \mathbf{Z} , $\tilde{\mathbf{Z}}$ be the variables that appear in \mathbf{Z} , but do not appear in \mathbf{X} or \mathbf{W} , and finally, let \mathbf{Q} be the common variables that appear in \mathbf{X} , \mathbf{W} , and \mathbf{Z} . Let $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\delta}}$, $\tilde{\boldsymbol{\gamma}}$, and $\boldsymbol{\alpha}$ be the vectors of coefficients on $\tilde{\mathbf{X}}$, $\tilde{\mathbf{W}}$, $\tilde{\mathbf{Z}}$, and \mathbf{Q} , respectively.

If these models were artificially nested, the result would be

$$\mathbf{y} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{W}}\tilde{\boldsymbol{\delta}} + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\gamma}} + \mathbf{Q}\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

The problem with this approach is that a test of $\tilde{\boldsymbol{\beta}} = \mathbf{0}$ or $\tilde{\boldsymbol{\delta}} = \mathbf{0}$ or $\tilde{\boldsymbol{\gamma}} = \mathbf{0}$ does not discriminate between the models because $\boldsymbol{\alpha}$, which is a mix of $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\delta}}$, and $\tilde{\boldsymbol{\gamma}}$, is not included in any of these tests. Instead of discriminating between the three models above, the likelihood ratio test in this circumstance discriminates between the three original models and a hybrid model comprising the null hypothesis and the alternatives. An artificially nested test cannot answer which, if any, of these models outperforms its rivals.

Other problems with the artificial nesting approach include low power (McAleer 1987), collinearity (Greene 2003), and the atheoretical nature of the nested model. Even if the models do not share variables, the artificial nesting approach is not recommended. No major econometrics text—Greene (2003), Kmenta (1986), Judge, Griffiths, Hill, Lutkepohl, & Lee (1985), Johnston & DiNardo (1997), Davidson & MacKinnon (1993), Mittelhammer, Judge, & Miller (2000), or Cameron & Trivedi (2005)—presents it as a viable option.

3 Specifications as Treatments

To understand how it is possible to conceive of model specifications as treatments, it is necessary to review the analysis of what statisticians call randomized complete block designs (RCBDs). In these situations, several related samples are used in an experiment to detect differences between three or more different treatments. The examples to follow do not come from political science for two important reasons. One, RCBDs are rarely used in political science, and two, the examples we have chosen are clearer than any political science example could be. (The use to which we will put RCBDs later in the paper is quite different from the use for which the tests were designed.)

Consider an experiment to compare four different animal diets (Zar 1999, 250). The question to be answered is whether the four diets have different effects on the body weight of the animals. Assume that there are 20 animals

to study; four sibling animals from each of five different sets of parents. Further, assume that each of four sibling animals from a particular set of parents is assigned randomly to eat each of the four diets. The body weights of each set of four animals form a block. This experiment is depicted in Table 2. The four diets constitute the treatments, the five sets of related offspring constitute the blocks, and each X_{ij} represents the body weight of an animal.

Table 2: Randomized Complete Block Design

Block	Diets			
	1	2	3	4
Siblings 1	X_{11}	X_{12}	X_{13}	X_{14}
Siblings 2	X_{21}	X_{22}	X_{23}	X_{24}
Siblings 3	X_{31}	X_{32}	X_{33}	X_{34}
Siblings 4	X_{41}	X_{42}	X_{43}	X_{44}
Siblings 5	X_{51}	X_{52}	X_{53}	X_{54}

The difference between the above experiment and a simple randomized design is that the four animals in each block are related (they are siblings). In other words, the four samples are not independent of one another. A block, then, is a generalization of the “paired” concept familiar from the paired t -test commonly taught in introductory statistics courses.

In another form of the design, a block can comprise a single unit. Conover (1999, 368) provides the following example:

Seven different men are used in a study of the effect of color schemes on work efficiency. Each man is considered to be a block

and spends some time in each of three rooms, each with its own type of color scheme. While in the room, each man performs a work task and is measured for work efficiency. The three rooms are the treatments.

This version of the randomized complete block design is depicted in Table 3. The three rooms constitute the treatments, the seven men form the blocks, and each X_{ij} represents the work output of a subject.

Table 3: Randomized Complete Block Design

Block	Rooms		
	1	2	3
Subject 1	X_{11}	X_{12}	X_{13}
Subject 2	X_{21}	X_{22}	X_{23}
Subject 3	X_{31}	X_{32}	X_{33}
Subject 4	X_{41}	X_{42}	X_{43}
Subject 5	X_{51}	X_{52}	X_{53}
Subject 6	X_{61}	X_{62}	X_{63}
Subject 7	X_{71}	X_{72}	X_{73}

This particular design can be problematic when there exist carryover or residual effects from one treatment to the next (Bradley 1968, 124). If the treatments were diets instead of room colors, for example, there would likely be large carryover effects. A man eating diet 1 might experience physiological changes that, when he switched to diet 2, would affect the efficacy of diet 2. Thus, separating out the effects of each diet would prove difficult.

When the treatments are room colors, on the other hand, there are likely to be no carryover or residual effects. It is difficult to imagine that the effect of

the color scheme of a room lingers once the subject has left the room. Having a single subject experience all three treatments, therefore, is not only not a problem, it is helpful in the sense that idiosyncracies that may exist within the blocks are controlled.⁴

It is this version of the randomized complete block design that is statistically identical to the application of different model specifications to the same set of observations. The application of a specification to a data set in no way changes the data set. Thus, there are no carryover effects. When this design is applied to competing models, the rival specifications are the treatments, and each observation, whether it be a voter, a state, a country, or even a militarized interstate dispute, is a block.⁵ The measured result of each treatment is the individual log-likelihood, which is the likelihood of that specification having produced that particular observation (generated when the maximum likelihood estimates are plugged into the likelihood equation for the parameter values). This design is depicted in Table 4; note the isomorphism with the design depicted in Table 3.

The point of seeing model specifications as treatments is that the tools for analyzing randomized complete block designs can be applied to the problem of discriminating between rival models. Under the assumptions of normality,

⁴If the tasks are related, it is possible that learning might occur, which is different from a carryover effect. Neither of these processes is an issue in the following discussion.

⁵The use of the term “treatment” here does not imply that the variable must be manipulable. Gibbons & Chakraborti (1992, 391) write, “the word treatment effect is used in a very general way and may not refer to a real treatment. It may refer to the effect of a condition or characteristic such as income level or race.”

Table 4: Specifications as Treatments

Specifications			
Block	1	2	3
1	$\log\text{Lik}_{11}$	$\log\text{Lik}_{12}$	$\log\text{Lik}_{13}$
2	$\log\text{Lik}_{21}$	$\log\text{Lik}_{22}$	$\log\text{Lik}_{23}$
3	$\log\text{Lik}_{31}$	$\log\text{Lik}_{32}$	$\log\text{Lik}_{33}$
\vdots	\vdots	\vdots	\vdots
n	$\log\text{Lik}_{n1}$	$\log\text{Lik}_{n2}$	$\log\text{Lik}_{n3}$

independence, and homoscedasticity, the correct procedure for analyzing this model is parametric analysis of variance. The next section discusses why parametric ANOVA fails in this case, and how a nonparametric procedure can be substituted in its place.

4 Procedures for Analyzing RCB Designs

In analysis of variance for randomized complete block designs, the total sum of squares is partitioned into three parts: the sum of squares for blocks, treatments, and error (Wackerly, Mendenhall, & Scheaffer 2002, 655). The model is generally written

$$X_{ij} = \mu + \beta_i + \theta_j + \epsilon_{ij},$$

for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, k$ (Gibbons & Chakraborti 1992, 384). The β_i are the row, or block, effects, and the θ_j are the column, or treat-

ment, effects. The error terms, ϵ_{ij} , are assumed to be independent, normally distributed random variables with mean zero and variance σ_ϵ^2 . The null hypothesis is that the column effects, or treatment effects, are equal,

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_k.$$

The alternative hypothesis is that at least one of the column, or treatment, effects is different,

$$H_1 : \theta_i \neq \theta_j \text{ for at least one } i \neq j.$$

Under the assumption of normality, the appropriate test statistic is distributed as an F .

The assumption of normally distributed errors is clearly violated, however, in the situation where model specifications are the treatments and the X_{ij} are individual log-likelihoods, which are always negative and left-skewed. Nonparametric ANOVA makes no assumption of normality, and a nonparametric procedure exists for the analysis of RCB designs: the Friedman test. It may appear odd at first glance to pair parametric likelihood-based models with a nonparametric model selection test. In actuality, combinations of the parametric and the nonparametric are common. For example, many of the techniques used to obtain standard errors for parametrically derived point estimates, such as the jackknife and the bootstrap, are nonparametric

(Efron 1981). Employing parametric models does not imply the necessary use of parametric tests.

The novelty of this approach to multiple model testing lies in the fact that the new procedures are actual tests, and thus provide probabilistic statements regarding model selection. Using tests for model selection is an approach that dates back to Vuong (1989) and contrasts sharply with model selection criteria such as AIC and BIC. Whereas model selection criteria always choose a “best” model—they provide no measure of the uncertainty regarding the choice—model selection tests allow no model to be chosen if the rival models are statistically equivalent. Model selection tests, therefore, provide substantial improvements over model selection criteria without, as discussed below, a substantial increase in the cost of calculation.

The next section describes the Friedman test.⁶ Although the test is widely available in the standard statistical software packages, it is treated here in some detail as political scientists are unlikely to be familiar with it.

4.1 The Friedman Test

To compute the Friedman test, replace the log-likelihoods, $\log\text{Lik}_{ij}$, in each block of Table 4 with their ranks from smallest to largest. The result is in Table 5.

⁶Quade’s (1979) test is a less well known alternative. See Iman, Hora, & Conover (1984) for a comparative analysis with the Friedman test.

Table 5: Friedman Test

Specifications			
Block	1	2	3
1	r_{11}	r_{12}	r_{13}
2	r_{21}	r_{22}	r_{23}
3	r_{31}	r_{32}	r_{33}
\vdots	\vdots	\vdots	\vdots
n	r_{n1}	r_{n2}	r_{n3}

If the effects of the treatments/specifications were equal, the average rank, $(k + 1)/2$, would be assigned to all three cells in the row. Let R_j be the sum of the ranks for the j th treatment, $R_j = \sum_{i=1}^n r_{ij}$, and \bar{R}_j be the average rank for the j th treatment, $\bar{R}_j = (\sum_{i=1}^n r_{ij})/n$. The expected value of the average rank for the j th treatment, $E(\bar{R}_j)$, under the null hypothesis of no difference, is then $(k + 1)/2$,

$$\begin{aligned} E_0(\bar{R}_j) &= E_0\left(\frac{1}{n}R_j\right) \\ &= \frac{1}{n}E_0\left(\sum_{i=1}^n r_{ij}\right) \\ &= \frac{k + 1}{2}. \end{aligned}$$

As Hollander & Wolfe (1999, 277) point out, the expected \bar{R}_j 's should be close to $(k + 1)/2$ under the null hypothesis. The test statistic is therefore the sum of the squared differences between the observed treatment average ranks and the expected value under the null,

$$S = \frac{12n}{k(k+1)} \sum_{j=1}^k \left(\bar{R}_j - \frac{k+1}{2} \right)^2 .$$

When the null hypothesis is false, the average treatment effects differ from $(k+1)/2$, and the squared term in S increases. The null is rejected, then, for large values of S .

Figure 1: Calculating the Friedman Test

1. Calculate the individual log-likelihoods for each model.
2. For each block/observation, rank the k log-likelihoods from smallest to largest.[†]
3. Average the ranks for each treatment/specification, \bar{R}_j .
4. Calculate the test statistic, S .
5. Reject H_0 if $S \geq s_\alpha$.

[†]Special procedures exist to deal with ties (see Hollander & Wolfe 1999, 273), but ties are highly unlikely when working with log-likelihoods.

Specific instructions on calculating the Friedman test are given in Figure 1. Selected values of s_α , the critical value, are in Bradley (1968), Gibbons & Chakraborti (1992), and Hollander & Wolfe (1999). Tables are rarely needed, however, as Steps 2-5 are performed automatically using the `friedman` command in Stata and the `friedman.test` command in *R*. For large samples, the critical value s_α can be approximated by a chi-square distribution with $k - 1$ degrees of freedom (Wackerly, Mendenhall, & Scheaffer 2002, 739).

The assumptions of the test are quite general. The n blocks/observations are

mutually independent (one block does not influence any other block), and the observations within each block can be ranked according to some criterion of interest (Conover 1999, 369).⁷ When a unit is treated as a block, it is assumed that there are no carryover or residual effects between treatments. The test is considered to be distribution-free because no assumption is made about the distribution of the data.

A question that remains is how much more efficient the Friedman test is than the corresponding normal theory test when the assumption of normality is violated. The asymptotic relative efficiency (see Davidson & MacKinnon 1993, 421), or A.R.E., of the Friedman test versus the normal theory F test is given by Bradley (1968, 125) as

$$\left(\frac{3}{\pi}\right) \left(\frac{k}{k+1}\right).$$

When there are three treatments ($k = 3$), the A.R.E. is 0.713, which means that it would take 40% more observations for the F -test to be as efficient as the Friedman test. The greater efficiency of the Friedman test slowly decreases as k increases toward infinity.⁸ When there are only two treatments, $k = 2$, the efficiency is 0.637, which is exactly equal to the efficiency of

⁷Although the Friedman test is fairly robust to violations of the independence assumption, the assumption should be checked, and we provide evidence of independence in the substantive example in Section 6. The independence of observation assumption underlies almost all inference in econometrics, and the assumption is accurate to the extent that the dependence is modelled in the statistical equation.

⁸The reason is that with more treatments, the ranks become finer grained (Bradley 1968, 125).

the paired sign test. This equivalence indicates that using the Friedman test to discriminate between multiple nonnested model specifications is a direct generalization of Clarke's procedure for discriminating between two rival nonnested models (Clarke 2003, Clarke 2007).

4.2 Friedman Test Example

An example makes the procedure clear. Assume that a researcher is comparing three rival model specifications on nine observations. Table 6 contains the individual log-likelihoods and associated ranks for the three specifications.

Table 6: Friedman Test Example

Observation	Log-Likelihoods(Ranks)		
	1	2	3
1	-0.4071(3)	-0.4196(2)	-0.6823(1)
2	-0.8705(3)	-1.1288(1)	-1.0509(2)
3	-0.3311(3)	-0.4376(1)	-0.4059(2)
4	-0.9632(3)	-1.0613(2)	-1.0709(1)
5	-0.2121(3)	-0.4175(2)	-0.4351(1)
6	-0.4928(1)	-0.4605(2)	-0.4402(3)
7	-0.3782(3)	-0.4799(1)	-0.4452(2)
8	-0.2936(3)	-0.4051(1)	-0.4037(2)
9	-0.2914(3)	-0.4027(2)	-0.4029(1)
\bar{R}_j	25	14	15
\bar{R}_j	2.78	1.56	1.67

Using the average ranks for each treatment given in Table 6, compute the test statistic,

$$\begin{aligned}
S &= \frac{12n}{k(k+1)} \sum_{j=1}^k \left(\bar{R}_j - \frac{k+1}{2} \right)^2 \\
&= \frac{12(9)}{3(3+1)} [(2.78 - 2)^2 + (1.56 - 2)^2 + (1.67 - 2)^2] \\
&= 8.2
\end{aligned}$$

Given $n = 9$ and $k = 3$, the p-value for this test from Table A.24 of Hollander & Wolfe (1999) is approximately 0.016. (The same value is returned by R 's `friedman.test` command.) The result indicates rejection of the null hypothesis of no differences between the three treatments. Having concluded that the treatments are different, it is necessary to determine which particular treatments led to the rejection of the null hypothesis, and this discussion is the subject of Section 5.

4.3 Adjusting the Test for Model Size

The Friedman test, when applied to model specifications, makes use of the individual log-likelihoods, which are sensitive to the dimensionality of the rival models. As the log-likelihoods increase with every additional estimated coefficient, a model specified with a large number of extraneous coefficients appears to fit better than a leaner model specified with only relevant coefficients. Thus, any model discrimination technique that makes use of log-likelihoods must include some kind of correction for model dimensionality.

Figure 2: Calculating the Adjusted Friedman Test

1. Calculate the individual log-likelihoods for each model.
2. Correct the individual log-likelihoods for each model by $[(p_j/2n) \ln n]$.
3. For each block/observation, rank the k log-likelihoods from smallest to largest.[†]
4. Average the ranks for each treatment/specification, \bar{R}_j .
5. Calculate the test statistic, S .
6. Reject H_0 if $S \geq s_\alpha$.

[†]Special procedures exist to deal with ties (see Hollander & Wolfe 1999, 273), but ties are highly unlikely when working with log-likelihoods.

Both the Vuong test and the Clarke distribution-free test use a correction that corresponds to either Akaike's (1973) information criteria or Schwarz's (1978) Bayesian information criteria. Although no particular correction can be justified, it is also true that which correction is used rarely makes a difference. As using the Friedman test to discriminate between nonnested model specification is a direct generalization of Clarke's distribution-free procedure, the same correction is applied. Therefore, the individual log-likelihoods for each specification are corrected by a factor of $[(p_j/2n) \ln n]$, where p_j is the number of estimated coefficients in the particular model, and then the test is applied. (See Vuong 1989 and Clarke 2007 for a more detailed justification for this correction.) The revised instructions for calculating the Friedman test after correcting for model dimensionality are in Figure 2.

Use of the correction factor does not change the properties of the test; it

remains unbiased and consistent. This fact can be seen by noting that the distribution of the Friedman test statistic depends solely on the number of different rank configurations, $(k!)^n$, which remains unaffected by the correction (Hollander & Wolfe 1999, 278).

The next section discusses how rival models can be ranked once the null hypothesis of no difference has been rejected.

5 Multiple Comparisons

The procedures discussed to this point test whether k model specifications have the same effects; that is, whether the distributions have the same location,

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k.$$

The alternative hypothesis is that at least one of the column, or treatment, effects is different,

$$H_1 : \theta_i \neq \theta_j \text{ for at least one } i \neq j.$$

When the null hypothesis is rejected, however, these tests do not identify which specifications are different or if one specification outperforms its rivals. All the researcher knows is that at least one specification is significantly different from the others. When comparing three model specifications, however,

it is necessary to distinguish between the following three possibilities,

$$s_1 : \theta_1 \neq \theta_2 = \theta_3;$$

$$s_2 : \theta_1 = \theta_2 \neq \theta_3;$$

$$s_3 : \theta_1 \neq \theta_2 \neq \theta_3.$$

This job is done by a multiple comparison procedure where the three specifications are compared in pairs to determine which of the above statements is true. Letting $R_j = \sum_{i=1}^n r_{ij}$, specification i is not equal to specification j if

$$|R_i - R_j| \geq r_\alpha,$$

where r_α is chosen such that the probability of at least one type I error is equal to α . That is, α applies to all $k(k-1)/2$ pairs (i, j) of specifications simultaneously. Selected values of r_α are in Hollander & Wolfe (1999, 708).⁹

When the number of blocks/observations is large, r_α can be approximated. The decision rule is then specification i is not equal to specification j if

$$|R_i - R_j| \geq q_\alpha \left[\frac{nk(k+1)}{12} \right]^{1/2},$$

⁹Some specialists in simultaneous inference (see Miller 1981, Hochberg & Tamhane 1987) would argue multiple comparisons can be used in place of ANOVA-type tests, instead of in conjunction with them. We, however, are following the more common advice of Conover (1999) and Zar (1999).

where q comes from the distribution of the range of k independent $N(0, 1)$ variables. Multiple comparisons for the Friedman test are automated in R using the `friedmanmc` command in the `pgirmess` library.

In the example given in Section 4.2, the null hypothesis of no difference between the specifications was rejected. The multiple comparison procedure is used to determine why the null was rejected. The three comparisons are $|R_1 - R_2| = 11$, $|R_1 - R_3| = 10$, $|R_2 - R_3| = 1$. The critical value for a 5% error rate is 10. Using the rule given above,

$$|R_i - R_j| \geq r_\alpha,$$

one can conclude that specification 1 is significantly different from specifications 2 and 3, and specifications 2 and 3 are not significantly different from one another. The same conclusion is reached using the large-sample approximation.

The following section provides a substantive example. The point of the example is not to insist that there is only one method of comparing models or that other indicators of fit should be ignored. Rather, the point is to demonstrate the usefulness of multiple model tests and to convince political scientists that this technique deserve a place in their methodological toolboxes.

6 Revisiting the Democratic Peace

Huth & Allee (2002), in an effort to explain the democratic peace, compare three rival models of the link between domestic institutions and foreign policy decisionmaking: the Political Accountability Model, the Political Norms Model, and the Political Affinity Model. In general, they find that the evidence supports the Accountability Model [286]. In a nonnested test of two of the three models, the Accountability Model and the Norms Model, Clarke (2007) found evidence in favor of the Norms Model. A nonnested test of all three competing models, however, was not possible until now.

In the Accountability Model, “competitive elections, independent legislative powers, and the threat of military coups are the sources of accountability for leadership decisions in foreign policy” (Huth & Allee 2002, 101). The model comprises four key assumptions: the critical goal of incumbent leaders is the retention of their office; political opponents challenge incumbents at strategic junctures; political accountability varies across different domestic political institutions; and the greater the political vulnerability of leaders, the more risk-averse leaders are in their foreign policy.

Huth & Allee (2002) operationalize the Accountability Model using six variables: a measure of how democratic the challenger and target are, whether the dispute is at a stalemate, whether the dispute is part of an enduring rivalry, whether ethnic co-nationals are involved in the dispute, whether the situation is one of high military risk or uncertainty, and the resolve of the

Table 7: The Political Accountability Model, $n = 374$

Variable	Challenger		Target	
	Coefficient	S.E.	Coefficient	S.E.
Challenger level of democracy	-0.004	(0.017)	-	-
Target level of democracy	-	-	0.006	(0.017)
Democracy*stalemate	-0.030	(0.023)	-0.008	(0.019)
Control for stalemate	-0.418	(0.192)	-0.421	(0.160)
Democracy*enduring rivalry	0.000	(0.018)	-0.016	(0.014)
Control for enduring rivalry	0.106	(0.172)	0.233	(0.159)
Democracy*ethnic ties	-0.014	(0.018)	0.005	(0.016)
Control for ethnic ties	0.400	(0.143)	0.053	(0.121)
Democracy*military risk	0.010	(0.022)	-0.026	(0.016)
Control for military risk	-0.141	(0.198)	0.066	(0.229)
Target resolve*target democracy	-0.036	(0.014)	-	-
Target signal of resolve	-0.052	(0.118)	-	-
Military balance	0.931	(0.301)	-0.062	(0.443)
Local balance of forces	0.440	(0.144)	0.026	(0.186)
Strategic value	0.559	(0.148)	0.334	(0.142)
Common security interests	-0.432	(0.184)	-0.128	(0.175)
Target other dispute	0.300	(0.169)	0.360	(0.162)
Challenger other dispute	0.351	(0.164)	0.135	(0.164)
Constant	-1.916	(0.265)	-1.237	(0.230)
ρ	0.956	(0.021)		
Log-Likelihood	-243.063			

target. Each of the last five variables is interacted with democracy in order to understand how the greater accountability of democratic leaders affects the decision to escalate a crisis.

In the Norms Model, “attention shifts to the principles that shape political elite beliefs about how to bargain and resolve political conflicts,” and leaders from democratic and non-democratic states have “different beliefs about the

Table 8: The Political Norms Model, $n = 374$

Variable	Challenger		Target	
	Coefficient	S.E.	Coefficient	S.E.
Strength of nonviolent norms	-0.009	(0.021)	-0.020	(0.018)
Nonviolent norms*mil. advantage	-0.017	(0.018)	-0.016	(0.014)
Nonviolent norms*stalemate	0.005	(0.036)	0.007	(0.043)
Control for stalemate	-0.345	(0.429)	-0.425	(0.509)
Military balance	1.201	(0.377)	0.057	(0.387)
Local balance of forces	0.535	(0.152)	0.085	(0.176)
Strategic value	0.485	(0.144)	0.353	(0.144)
Common security interests	-0.404	(0.190)	-0.136	(0.176)
Target other dispute	0.404	(0.167)	0.408	(0.163)
Challenger other dispute	0.273	(0.166)	0.099	(0.166)
Constant	-1.714	(0.328)	-0.917	(0.300)
ρ	0.924	(0.027)		
Log-Likelihood	-255.782			

acceptability of compromising with and coercing political adversaries” (Huth & Allee 2002, 101). The model comprises three key assumptions: norms influence decisions made by political actors in political conflict; domestic political institutions structure political conflict; and the bargaining strategies used by leaders in international disputes are influenced by the norms of bargaining those same leaders use with domestic political opponents.

Huth & Allee (2002) operationalize the Norms Model with four variables: the strength of nonviolent norms in the state, whether the dispute is at a stalemate, nonviolent norms interacted with stalemate, and nonviolent norms interacted with a measure of whether the state has a military advantage.

In the Affinity Model, “common political institutions and ideologies between

Table 9: The Political Affinity Model, $n = 374$

Variable	Challenger		Target	
	Coefficient	S.E.	Coefficient	S.E.
Political similarity	0.333	(0.270)	0.219	(0.251)
Recent change to pol. similarity	-1.629	(0.500)	-0.637	(0.474)
Military balance	0.920	(0.299)	-0.293	(0.287)
Local balance of forces	0.571	(0.148)	0.081	(0.178)
Strategic value	0.545	(0.141)	0.342	(0.139)
Common security interests	-0.478	(0.193)	-0.131	(0.180)
Target other dispute	0.354	(0.169)	0.397	(0.159)
Challenger other dispute	0.293	(0.170)	0.150	(0.167)
Constant	-1.763	(0.208)	-1.105	(0.211)
ρ	0.925	(0.027)		
Log-Likelihood	-256.367			

states produce shared political interests among those states' incumbent elites regarding whether preservation or change in the political status quo is desirable" (Huth & Allee 2002, 124). Leaders with similar institutions and ideologies see each other as allies and adopt more cooperative foreign policies toward one another. The model rests on three assumption: leaders attempt to use foreign policy to secure their hold power, in-groups and out-groups are a central feature of political identity formation, and that political conflict is greater between groups than within groups.

Huth & Allee (2002) operationalize the Affinity Model with two variables: whether the challenger and target countries share the same regime type, and whether the challenger and target countries have only recently become politically similar (five or fewer years).

The three models are completed by pairing the predictors listed above with a set of straightforward realist variables comprising military balance, local balance of forces advantage, the strategic value of the territory, alliance behavior, and whether the target or challenger are involved in another militarized dispute.¹⁰ This situation is precisely the one described in Section 2 where the models are nonnested, but share a subset of variables. The models are tested on 374 territorial disputes where the challenger has opted for military pressure over calling for negotiations. The challenger and the target both choose to either escalate the dispute or not (the dependent variable); consequently, the models are estimated by bivariate probit. The results in Tables 7, 8, and Tables 9 replicate the results of Huth and Allee's Tables 9.4 [240], 9.13 [256], and 9.19 [268], respectively.

The Friedman test statistic is 17.87 on two degrees of freedom, and the corresponding p-value is approximately zero.¹¹ It is reasonable to conclude that the models are different based on this result. Table 10 displays the three multiple comparisons for the three models. The critical value, r_α , for a significance level of 0.05 is 65.47. The sum of the ranks, R_j , for each of the specifications are: $R_{Accountability} = 653$, $R_{Norms} = 726$, and $R_{Affinity} = 865$.

The first thing to notice about the results is the agreement with the earlier study done by Clarke (2007); the Norms Model outperforms the Accountabil-

¹⁰See Huth & Allee (2002) for information on how these variables are operationalized.

¹¹Autocorrelation functions and partial autocorrelation functions calculated for the 3 sets of individual log-likelihoods show no statistically significant dependence. Graphs are available from the author.

Table 10: Multiple Comparisons, $r_{0.05} = 65.47$

Comparison	Observed Difference	Significant
1 against 2	73	Yes
1 against 3	212	Yes
2 against 3	139	Yes

ity Model. This finding corroborates earlier work on the effect of domestic politics on foreign policy decisionmaking (Maoz & Russett 1993). Even more interesting, though, is the result that the Affinity Model appears to outperform both the Accountability Model and the Norms Model. This result contradicts, in part, the conclusions of Huth and Allee, who write that the “Affinity Model produced the weakest results” [283].

This result may seem counterintuitive given that Accountability Model actually has the largest log-likelihood. The test chooses the Affinity Model, however, because of the correction factor. The Accountability Model is significantly more complex than the Affinity Model, but the improved fit does not compensate for the loss of parsimony. Use of information criteria such as BIC (Bayesian information criteria) and consistent AIC (consistent Akaike information criteria) leads to the same decision, although without the probabilistic statement of uncertainty that is the hallmark of hypothesis tests.

Though informative, these results are not definitive. Huth and Allee consider three stages in the evolution of international disputes: the decision to challenge the status quo, the decision to offer concessions in negotiations, and the decision to escalate the conflict with military force. The model comparison

presented here concerns only the decision to escalate the conflict. Thus, it is possible that the Affinity Model outperforms its rivals only at this stage. Even if this were the case, the finding is an important one and should serve to temper Huth and Allee's conclusions.

7 Conclusion

Theories must be tested against other theories, not simply against null hypotheses. The challenge of testing a theory against its rivals is that the competing theories are often nonnested, and nonnested models require special procedures. Until now, such procedures existed solely for the pairwise testing of nonnested models. This paper proposes the first classical solution to the problem of testing three or more competing nonnested models. By seeing rival model specifications as treatments applied to a set of observations, the tools for analyzing randomized complete block designs can be used to discriminate between them. The result is a set of procedures that are simple, possess good statistical properties, take advantage of existing software, and provide researchers the ability to determine whether a particular model outperforms its rivals. The usefulness of the new test is demonstrated with an application to the democratic peace, and the results indicate that by failing to test rival theories against one another appropriately, we may be making serious errors.

References

- Akaike, H. 1973. "Information Theory and an Extension of the Likelihood Ratio Principle." In *Second International Symposium of Information Theory*, ed. B.N. Petrov & F. Csaki. Minnesota Studies in the Philosophy of Science, Budapest pp. 267–281.
- Bradley, James V. 1968. *Distribution-Free Statistical Tests*. New Jersey: Prentice-Hall.
- Cameron, A. Colin, & Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.
- Chipman, Hugh, Edward I. George, & Robert E. McCulloch. 2001. "The Practical Implementation of Bayesian Model Selection." In *Model Selection*, ed. P. Lahiri. Vol. 38 of *Institute of Mathematical Statistics Lecture Notes* Beachwood Ohio: Institute of Mathematical Statistics.
- Clarke, Kevin A. 2003. "Nonparametric Model Discrimination in International Relations." *Journal of Conflict Resolution* 47 (February): 72-93.
- Clarke, Kevin A. 2007. "A Simple Distribution-Free Test for Nonnested Hypotheses." *Political Analysis* 15 (3).
- Conover, W.J. 1999. *Practical Nonparametric Statistics*. 2nd ed. New York: John Wiley and Sons.

- Cox, David R. 1961. "Tests of separate families of hypotheses." *Proceedings of the Fourth Berkeley Symposium I*: 105-123.
- Davidson, Russell, & James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
- Efron, Bradley. 1981. "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods." *Biometrika* 68 (3): 589-599.
- Efron, Bradley. 1986. "Why Isn't Everyone a Bayesian?" *The American Statistician* 40: 1-5.
- Feyerabend, Paul. 1968. "How to be a good empiricist—a plea for tolerance in matters epistemological." In *The Philosophy of Science*, ed. P.H. Niddich. London: Oxford University Press.
- George, Edward I. 2000. "The Variable Selection Problem." *Journal of the American Statistical Association* 95 (December): 1304-1308.
- Gibbons, Jean Dickinson, & Subhabrata Chakraborti. 1992. *Nonparametric Statistical Inference*. 3rd ed. New York: Marcel Dekker, Inc.
- Greene, William H. 2003. *Econometric Analysis*. 5 ed. New Jersey: Prentice Hall.
- Hochberg, Yosef, & Ajit Tamhane. 1987. *Multiple Comparison Procedures*. New York: John Wiley and Sons.

- Hollander, Myles, & Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. 2 ed. New York: John Wiley and Sons.
- Huth, Paul K., & Todd L. Allee. 2002. *The Democratic Peace and Territorial Conflict in the Twentieth Century*. New York: Cambridge University Press.
- Iman, Ronald L., Stephen C. Hora, & William J. Conover. 1984. "Comparison of Asymptotically Distribution-Free Procedures for the Analysis of Complete Blocks." *Journal of the American Statistical Association* 79 (September): 674-685.
- Johnston, Jack, & John DiNardo. 1997. *Econometric Methods*. 4 ed. New York: McGraw-Hill.
- Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, & Tsoung-Chao Lee. 1985. *The Theory and Practice of Econometrics*. 2 ed. New York: John Wiley and Sons.
- Kmenta, Jan. 1986. *Elements of Econometrics*. 2 ed. New York: Macmillan Publishing Company.
- Lakatos, Imre. 1981. "History of Science and Its Rational Reconstructions." In *Scientific Revolutions*, ed. Ian Hacking. London: Oxford University Press.

- Lawrence, Eric D., Forrest Maltzman, & Steven S. Smith. 2006. "Who Wins? Party Effect in Legislative Voting." *Legislative Studies Quarterly* 31 (February): 33-69.
- Maoz, Zeev, & Bruce Russett. 1993. "Normative and Structural Causes of Democratic Peace, 1946-1986." *American Political Science Review* 87 (September): 624-638.
- Martin, Andrew D., & Kevin M. Quinn. 2006. "Applied Bayesian Inference in R using MCMCpack." *R News* 6 (March): 2-7.
- McAleer, Michael. 1987. "Specification Tests for Separate Models: A Survey." In *Specification Analysis in the Linear Model*, ed. M.L. King & D.E.A. Giles. London: Routledge and Kegan Paul.
- Miller, Rupert G. 1981. *Simultaneous Statistical Inference*. New York: Springer-Verlag.
- Mittelhammer, Ron C., George G. Judge, & Douglas J. Miller. 2000. *Econometric Foundations*. New York: Cambridge University Press.
- Popper, Karl. 1965. *Conjectures and Refutations*. New York: Harper and Row.
- Quade, Dana. 1979. "Using Weighted Rankings in the Analysis of Complete Blocks with Additive Block Effects." *Journal of the American Statistical Association* 74 (September): 680-683.

- Rosenbaum, Paul R. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science* 14 (August): 259-278.
- Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6: 461-464.
- Tukey, John W. 1977. "Some Thoughts on Clinical Trials, Especially Problems of Multiplicity." *Science* 198 (November): 4318.
- Volden, Craig, & Clifford J. Carrubba. 2004. "The Formation of Oversized Coalitions in Parliamentary Democracies." *American Journal of Political Science* 48 (July): 521-537.
- Vuong, Quang. 1989. "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica* 57 (March): 307-333.
- Wackerly, Dennis D., William Mendenhall, & Richard L. Scheaffer. 2002. *Mathematical Statistics with Applications*. 6th ed. Pacific Grove, CA: Duxbury.
- Zar, Jerrold H. 1999. *Biostatistical Analysis*. 4th ed. New Jersey: Prentice Hall.