MONASH

2/94

M O N A S H

U N I V E R S I T Y

AUSTRALIA

TESTING FOR INDEPENDENCE OF IRRELEVANT ALTERNATIVES: SOME EMPIRICAL RESULTS

Tim^othy^ R.L. Fry and Mark N. Harris

Working Paper No. 2/94

April 1994

DEPARTMENT OF ECONOMETRICS

ISSN 1032-3813

ISBN 0 7326 0391 9

TESTING FOR INDEPENDENCE OF IRRELEVANT ALTERNATIVES: SOME EMPIRICAL RESULTS

Tim R.L. Fry and Mark N. Harris

Working Paper No. 2/94

April 1994

DEPARTMENT OF ECONOMETRICS

MONASH UNIVERSITY, CLAYTON, VICTORIA 3168, AUSTRALIA.

Testing for Independence of Irrelevant Alternatives:

Some Empirical Results.

Tim R.L. Fry and Mark N. Harris[1]

Department of Econometrics
Monash University
Clayton, Victoria 3168
Australia.

Abstract: We estimate a multinomial Logit (MNL) model of U.K. Magistrates' Courts sentencing using a data set collected by the National Association for the Care and Resettlement of Offenders (NACRO) and test the independence of irrelevant alternatives (IIA) property using six tests. Conducting the tests with the appropriate asymptotic critical values we find that the acceptance or rejection of IIA depends both upon which test and which variant of a given test is used. The same tests are then performed using empirical critical values obtained by simulation and the resultant inferences compared. Our results show that empirical workers should exercise care when testing for IIA.

Keywords: Multinomial Logit, Independence of Irrelevant Alternatives, Hypothesis Testing.

J.E.L. Classification: C25.

---

[1]Part of this work was supported by a grant from the Australian Research Council. We are grateful to NACRO for allowing us access to the data. The views expressed in this paper are not necessarily those of NACRO.

1. Introduction.

This paper concerns the testing of a behavioral assumption implicit in the use of Logit models for polychotomous choice data. The Independence of Irrelevant Alternatives (IIA) property implies that the choice between two outcomes in a choice set depends solely upon the characteristics of the outcomes being compared and not upon the characteristics, or indeed, the existence, of any other outcomes in the choice set.

The assumption of the IIA property facilitates model estimation. However, it may be unduly restrictive in terms of individual behavior, and inconsistent with the available survey data on polychotomous choice situations. Thus it is widely acknowledged that when using Logit models for discrete choice behavior, researchers should test for the IIA property (see, for example, Fry *et al* (1993), Zhang and Hoffman (1993)). Many tests for the IIA property have been proposed in the literature. This paper considers six tests, which are applied to a previously analyzed data set on court sentencing patterns. The purpose of the paper is to compare the outcomes of the tests to ascertain whether the same inference is made across different tests and also within the various versions of the same test. This inference is conducted using both asymptotic critical values and empirical critical values, the latter obtained from a simulation experiment. Finally, the results are compared across the two procedures.

The plan of the rest of this paper is as follows. In the next section we discuss the specification of Logit models of discrete choice behavior and testing for the IIA property. Section three contains the results of testing for the IIA property on the court sentencing data and finally section four contains some concluding remarks.

2. Testing for Independence of Irrelevant Alternatives.

In this section, the specification of Logit models of discrete choice behavior is discussed, along with the associated IIA property. The six tests for this property that are used in this paper are then described. As is common in the literature (see, for example, Fry *et al* (1993)) a random utility maximization (RUM) model is used. The (indirect) utility function is given by:

$$U_{ij} = V_{ij}(\mathbf{Z}_{ij}, \mathbf{X}_i) + \varepsilon_{ij}, \qquad i = 1, ..., n; \;\; j = 1, ..., J, \tag{1}$$

where U_{ij} is the utility individual i derives from choosing alternative j. This is comprised of two components, V_{ij} and ε_{ij}. The former is a deterministic component which depends upon characteristics of the individual $\mathbf{X}_i$ and variables which vary across both individuals and alternatives, $\mathbf{Z}_{ij}$. The latter, ε_{ij}, is the random component which represents unobservable factors. Typically the functional form for the V_{ij} in equation (1) is assumed to be linear, thus:

$$U_{ij} = \mathbf{Z}_{ij}'\alpha + \mathbf{X}_i'\beta_j + \varepsilon_{ij} = \mathbf{W}_{ij}'\delta + \varepsilon_{ij}, \qquad i = 1, ..., n; \;\; j = 1, ..., J. \tag{2}$$

If it is assumed that the ε_{ij}'s are independent and identically distributed as Extreme Value then a Logit model specification results with associated selection probabilities, P_{ij} = P(individual i selects alternative j), given by:

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{k=1}^{J}\exp(V_{ik})} \qquad i = 1, ..., n; \;\; j = 1, ..., \;\; J.$$

The IIA property states that the odds of choosing alternative *j* over alternative *k* ($k \neq j$), P_{ij}/P_{ik}, are independent of all other alternatives and of the number of alternatives in the choice set. Thus in this instance,

$$\frac{P_{ij}}{P_{ik}} = \frac{\exp(V_{ij})}{\exp(V_{ik})},$$

which, using the form for the V_{ij}'s as given in equation (2) above, will satisfy the definition of IIA. Although this property implies strong restrictions on behavior, its use leads to considerable advantages in model specification, estimation and forecasting. Given the frequent use of such models to explain choice behavior, it is important to test for the existence of the IIA property.

Probably the most widely used test for the IIA property is the Hausman-McFadden (HM) test (Hausman and McFadden (1984)). The HM test statistic is based upon the fact that if IIA holds, the model's structure will be invariant to whether the parameter estimates are obtained from the full choice set C or from a restricted subset, D, of this choice set. That is, consistent estimates of δ in (2) are obtained by maximizing the log-likelihood of the Logit model either over the full choice set C, or over the subset D, (yielding $\hat{\delta}_C$ and $\hat{\delta}_D$ respectively). The test is a Hausman specification test (Hausman (1978)) with two estimators being employed. One of these, $\hat{\delta}_D$, is both consistent and efficient under the null hypothesis of IIA, but inconsistent under the alternative that IIA does not hold. The other estimator, $\hat{\delta}_C$, is consistent under *both* the null and alternative but inefficient under the alternative. The HM test statistic is given by:

$$HM = (\hat{\delta}_D - \hat{\delta}_C)'\Omega^{-}(\hat{\delta}_D - \hat{\delta}_C) = \hat{q}'\Omega^{-}\hat{q},$$

where Ω^{-} is the generalized inverse of the asymptotic covariance matrix of $\hat{q}$. Asymptotically the test follows a chi-squared distribution with degrees of freedom equal to the rank of Ω^{-}. Hausman and McFadden show that Ω^{-} is asymptotically equivalent to $(\Omega_D - \Omega_C)^{-}$, the generalized inverse of the difference of the asymptotic covariance matrices of $\hat{\delta}_D$ and $\hat{\delta}_C$ respectively. This is the version of the HM test that we use in this paper.

The HM test can conveniently be termed a "choice set partitioned test" in that it involves partitioning the full choice set. In this paper three other so-called 'choice set partitioned' tests are also considered: the McFadden-Train-Tye (MTT) test (McFadden, Train and Tye (1981)), the Horowitz (H) test (Horowitz (1981)) and the Small-Hsiao (SH) test (Small and Hsiao (1985)). All of these tests are variants of the likelihood ratio test and are based upon the same basic premise. That is, if the IIA property is valid, then we can base tests upon the difference between the maximized log-likelihood of the model estimated freely over a restricted subset, D, and the same log-likelihood function evaluated at the parameter estimates obtained from maximization over the full choice set, C.

Specifically, the MTT test involves estimation of the Logit model by maximum likelihood on the full choice set C thus obtaining parameter estimates, $\hat{\delta}_C$. The Logit model is then estimated using data over the subset D, yielding $\hat{\delta}_D$. The log-likelihood for this 'restricted' (in the sense of a restricted choice set) estimation is labelled $\log L_1$ and its maximized value is $\log L_1(\hat{\delta}_D)$. The MTT test then compares $\log L_1(\hat{\delta}_D)$ with $\log L_1$ evaluated at the full choice set estimates $\hat{\delta}_C$ (i.e. with $\log L_1(\hat{\delta}_C)$). Thus the MTT test statistic is given by:

$$\text{MTT} = -2(\log L_1(\hat{\delta}_C) - \log L_1(\hat{\delta}_D)).$$

This test statistic has an asymptotic chi-squared distribution with degrees of freedom equal to the dimension of $\hat{\delta}_D$.

However, due to the use of overlapping estimation samples, a bias towards acceptance of the null hypothesis of IIA arises in the MTT test. To avoid this bias, Horowitz (1981) suggests randomly dividing the sample into two asymptotically equal parts A and B, with respective sample sizes n_A and n_B. These two samples are subsequently used in the construction of a test statistic. The Horowitz (H) test statistic is then constructed as follows. Firstly, the Logit model is estimated by maximum likelihood for data over the full choice set C in the sub-sample of n_A observations. This yields the estimates $\hat{\delta}_C^A$. The sub-sample A is now discarded. The Logit model is then estimated by maximum likelihood using data over the subset D, in the sub-sample of n_B observations. This estimation yields parameter estimates, $\hat{\delta}_D^B$, with associated maximized log-likelihood of $log\ L_1(\hat{\delta}_D^B)$. The H test statistic is calculated as:

$$H = -2(log\ L_1(\hat{\delta}_C^A) - log\ L_1(\hat{\delta}_D^B)),$$

where $log\ L_1(\hat{\delta}_C^A)$ is $log\ L_1$ evaluated at the full choice set estimates, $\hat{\delta}_C^A$, obtained from sub-sample A. This test statistic follows an asymptotic chi-squared distribution with degrees of freedom equal to the dimension of $\hat{\delta}_D^B$.

Small and Hsiao (1985) show that the use of independent samples causes the H test to be biased towards rejecting the null hypothesis of IIA. In their paper they propose a test (SH) that combines the MTT and H test procedures such that the resulting test is free of any bias. Again, the sample is randomly divided into two asymptotically equal parts A and B with respective sample sizes, n_A and n_B. The Logit model is then estimated by maximum likelihood in each sub-sample over the data for the full choice set C. These procedures yield estimates $\hat{\delta}_C^A$ and $\hat{\delta}_C^B$ which are then combined in a weighted average:

$$\hat{\delta}_C^{AB} = (1/\sqrt{2})\hat{\delta}_C^A + (1 - 1/\sqrt{2})\hat{\delta}_C^B.$$

The first sub-sample A is once more discarded and the Logit model is estimated by maximum likelihood in sub-sample B for data over the subset D. This yields parameter estimates, $\hat{\delta}_D^B$, and associated maximized log-likelihood $log\ L_1(\hat{\delta}_D^B)$. The SH test statistic is then calculated as:

$$SH = -2(log\ L_1(\hat{\delta}_C^{AB}) - log\ L_1(\hat{\delta}_D^B)),$$

where $log\ L_1(\hat{\delta}_C^{AB})$ is $log\ L_1$ evaluated at the weighted average of the full choice set estimates, $\hat{\delta}_C^{AB}$, defined above. Again, this test statistic follows an asymptotic chi-squared distribution with degrees of freedom equal to the dimension of $\hat{\delta}_D^B$.

Another class of tests for the IIA property involves the specification of an alternative model which does not embody IIA. Typically such models are generalizations of the Logit model and IIA is tested using conventional tests for parameter restrictions. Such a generalization is the DOGIT model of Gaudry and Dagenais (1979). The selection probabilities for the DOGIT model are given by:

$$P_{ij} = \frac{\exp(V_{ij}) + \theta_j \sum_{k=1}^{J} \exp(V_{ik})}{\left(1 + \sum_{k=1}^{J} \theta_k\right) \sum_{k=1}^{J} \exp(V_{ik})} \qquad i = 1, ..., n; \;\; j = 1, ..., J.$$

The DOGIT model is convenient in that if certain parameters (the θ's) are equal to zero it collapses to the Logit model. As the Logit model is nested within the DOGIT model then a test for the Logit, and hence for IIA, can be carried out by testing the appropriate parameter restriction using a Wald, likelihood ratio or Lagrange multiplier (score) test procedure.

From a practical viewpoint, the most appealing procedure is that of the score test, as it only involves estimation under the null hypothesis. That is, estimation of the Logit model. Tse (1987) derives the score (LM) test of the IIA property using the DOGIT model as his alternative (non-IIA) specification. The LM statistic however ignores the inherent one-sided nature of the hypothesis test in this case. In an earlier paper (Fry and Harris (1993)) we derive an appropriate one-sided test based upon the locally most mean powerful (LMMP) test procedure. This LMMP test statistic turns out to be nothing more than a standardized sum of scores and is thus straightforward to use in testing the Logit against the DOGIT specification. Therefore, as alternatives to the 'choice set partitioned tests' described above, the LM and LMMP tests against the DOGIT specification are also used as tests for the IIA property.

3. An Empirical Example.

In this section we take a 'real life' data set and test for the IIA property using the six tests above (HM, MTT, H, SH, LM and LMMP). The data concerns the sentencing of male offenders in English Magistrates Courts and is a subset of that used in Crichton and Fry (1992)[2]. For this data, there are six outcomes: Discharged, Probation, Community Service Order, Suspended Imprisonment, Immediate Imprisonment and Fined. Table 1 shows the number of offenders in each category in our full choice set (i.e. the set C). It is clear from this table that one outcome (Fined) dominates and one (Suspended Imprisonment) is selected relatively less frequently than the others. Such a split across outcomes is not uncommon in empirical work.

We estimate the Logit model and conduct the six tests for the IIA property as outlined above[3]. For the 'choice set partitioned' tests there are a large number of subsets, D, which could be formed from the complete set, C. We choose only to consider six choices of D. That is, the subsets formed by deleting each single outcome in turn. The tests are conducted using both asymptotic critical values, using the appropriate limiting distribution, and then using simulated 'empirical' critical values. The resultant inference is then compared.

Table 2 contains the calculated test statistic values. For the 'choice set partitioned' tests several points should be made. The values for the HM test differ quite markedly. The subset formed by dropping the Immediate Imprisonment outcome leads to a rejection of the null hypothesis of IIA. The negative value for the subset formed by dropping the Fined outcome arises as a consequence of the way in which we estimate Ω^{*} and is taken as indicating a non-rejection of the null

[2]That is to reduce computing times in our simulation work we use three of the six courts from their data set.
[3]The results of the estimation of the Logit model over the full data set and some summary statistics may be found in the Appendix.

(see footnote 4, page 1226 of Hausman and McFadden (1984)). Thus for the HM test one of the calculated test statistics rejects the null and the other five do not reject the null.

Both the MTT and the H tests reflect the theoretical biases described above. That is the MTT test is biased towards *not* rejecting the null and the H test is biased towards rejecting the null. Thus the calculated MTT test statistic values are all "small" and the calculated H test statistic values are all "large". These calculated values lead to six rejections of the null for the H test and six decisions not to reject the null for the MTT test. It is interesting to note that the theoretical biases of the MTT and H tests were also confirmed by previous Monte Carlo work (Fry and Harris (1993)). The SH test, on the other hand, corrects for these known biases and the calculated values lie between those of the MTT and H tests. Using the SH test we do not reject the null for five of the subsets but reject it for the subset formed by deleting Suspended Imprisonment which happens to be the smallest category.

Therefore, the 'choice set partitioned' tests lead to conflicting evidence concerning the null hypothesis of IIA. It possible for us to obtain conflicting inference not only across tests, but also within a given test. Such results are potentially worrying for empirical workers wishing to test for IIA. In other words, their conclusion concerning the validity of the IIA property (and hence of their model specification) appears to be dependant upon their choice of test statistic. It could be argued that neither the MTT test nor the H test should be considered because of their known biases, but we should note that a comparison of the HM and SH tests also reveals conflicting inference. Thus the problem is not just one of known bias.

We also computed the LM and LMMP tests for the Logit specification against the

DOGIT specification and in both cases find that the null hypothesis of the Logit model cannot be rejected. Thus these tests appear to support the null hypothesis of IIA holding. Unfortunately, earlier work (Fry and Harris (1993)) has indicated that these two tests are undersized in finite samples and thus potentially biased towards acceptance of the null Logit specification.

Thus, at this juncture, the researcher will be unsure as to whether they can accept the null hypothesis of IIA and therefore their Logit specification. Since earlier work (Fry and Harris (1993)) has shown evidence of the six tests for the IIA property used in this paper being incorrectly sized, inference using asymptotic critical values may be inappropriate. Therefore, an alternative approach to inference is the following simulation procedure.

Under the null that IIA holds, the correct specification is a Logit model. Thus we estimate a Logit model over the full choice set for the whole data set and use the estimated coefficients, $\hat{\delta}_C$, as "true" in a simulation experiment. Our simulation then involves using $\hat{\delta}_C$ in the RUM model given in (2) and adding to $W_{ij}'\hat{\delta}_C$ a random drawing from an Extreme Value distribution. We then map the resultant U_{ij}'s into an "observable variable" by finding the maximum U_{ij} value. The Logit model is estimated using this simulated data and the six tests computed. This set up is then replicated 1000 times and the empirical size performance assessed. The procedure also allows us to find the simulated empirical critical values for the test statistics. That is, the value of a given test statistic beyond which α% of all our calculated test statistics lie.

Table 3 contains our results on empirical size performance for a nominal size of 5%. These clearly show up the known biases in the MTT and H test statistics. Additionally they show that the SH test is oversized and the HM test is reasonably well sized. The simulations also indicate the extent to which the HM test can be affected by the anomaly of a negative calculated test statistic value. For example, 42.4% of calculated test statistics when the Discharged outcome is dropped to form the subset D, are negative (full details may be found in the Appendix). The results for the HM and SH tests somewhat contradict earlier findings (Fry and Harris (1993)), these found that, at the 5% level, the SH test was invariably correctly sized[4], whereas the HM test had fairly erratic size properties. It should be noted, however, that the earlier results were from a Monte Carlo experiment using a generated "X matrix" and that this may partly explain the apparent discrepancies between the two sets of size results (see Mátyás and Harris (1993) for a discussion of the impact of the choice of "X matrix" in Monte Carlo work). Once again, our results illustrate that the LM and LMMP tests appear to be severely undersized, concurring with earlier evidence.

One way to take account of differing size performance is to use simulated empirical critical values in our test procedures. These empirical critical values are found in Table 4. Conducting our inference using the empirical critical values leads to only two rejections - from a total of 26 test statistics - of the null hypothesis of IIA. These two rejections are HM, when Immediate Imprisonment is dropped to form the subset D, and MTT, when Fined is dropped to form the subset D. These results give a clearer picture, than that gained using asymptotic critical values, of the validity of the IIA property for the court sentencing problem. That is, that IIA and hence a Logit specification is likely to be acceptable.

4. Conclusions.

In this paper we have taken six tests for the Independence of Irrelevant Alternatives (IIA) property of Logit models for polychotomous choice situations and applied them to a real data set on court sentencing patterns. As is traditional in empirical work we conduct our inference using appropriate

[4]If the SH test was not correctly sized it was oversized.

asymptotic critical values. We compare the test results both across the different tests and within a given test across the variants of that test. Additionally, we carry out a Monte Carlo simulation experiment to assess the empirical size performance of the test statistics. This simulation study also gives us estimated 'empirical' critical values to use in our inference concerning the validity of the IIA property. Finally, the results of our inference using 'empirical' critical values is compared with that obtained using the asymptotic critical values.

Inference based upon asymptotic critical values revealed known biases. That is that the MTT test is biased towards the null (giving "small" calculated values) and that the H test is biased towards rejection of the null (giving "large" calculated values). However, even if we were to ignore the results of these two test procedures, we still find conflicting inference concerning the null of IIA both across tests and within a given test across its variants. Such conflicts imply that there is some uncertainty facing an empirical worker. That is, the decision made on the validity of the IIA property varies according to which test - or even which variant of a given test - is used. Thus it is possible for two empirical workers to come up with different answers using the same data set.

Our simulation study, based upon the model under the null (the Logit model), enabled us to investigate the size properties of the tests. These both confirmed and contradicted previous results in Fry and Harris (1993). The known biases for the MTT, H, LM and LMMP tests were confirmed. On the other hand, the SH test had worse size properties and the HM test better size properties with this sentencing data set. Using the 'empirical' critical values generated in our simulation yielded a more consistent inference across tests and within given tests. In this instance only 2 rejections of the null of IIA were found from 26 test statistics. From this we conclude that there is stronger evidence of the validity of the IIA property (or equivalently the Logit model).

Our results indicate that when testing for the IIA property we should exercise care in the choice of test statistic as conflicting inference can arise from differing choices of test statistic. This conflict appears most marked when inference is conducted using asymptotic critical values. It is likely that part of the problem arises through the poor size properties of the asymptotic procedures. Thus we suggest that, in practice, researchers use more than one test statistic to confirm whether any conflict of inference arises. Furthermore, where possible, we would recommend that a simulation experiment be conducted to obtain 'empirical' (size corrected) critical values for use in inference concerning the IIA property.

References.

Crichton, N.J. and T.R.L. Fry (1992), "An Analysis of the Effect of an Offender's Employment Status on the Type of Sentence Chosen by the Magistrate", in P.G.M. van der Heijden *et al* (eds.), *Statistical Modelling*, Elsevier Science, Amsterdam.

Fry, T.R.L., Brooks, R.D., Comley, B.R. and J. Zhang (1993), "Economic Motivations for Limited Dependent and Qualitative Variable Models", *Economic Record*, **69**, 193-205.

Fry, T.R.L. and M.N. Harris (1993), "A Monte Carlo Study of Tests for the Independence of Irrelevant Alternatives", Working Paper 8/93, Department of Econometrics, Monash University.

Gaudry, M.J.I. and M.G. Dagenais (1979), "The Dogit Model", *Transportation Research*, **13B**, 105-112.

Hausman, J.A. (1978), "Specification Tests in Econometrics", *Econometrica*, **46**, 1251-1271.

Hausman, J.A. and D. McFadden (1984), "Specification Tests for the Multinomial Logit Model", *Econometrica*, **52**, 1219-1240.

Horowitz, J. (1981), "Identification and Diagnosis of Specification Errors in the Multinomial Logit Model", *Transportation Research*, **15B**, 345-360.

McFadden, D., Train, K. and W. Tye (1981), "An Application of Diagnostic Tests for the Independence of Irrelevant Alternatives Property of the Multinomial Logit Model", *Transportation Research Record*, **637**, 39-46.

Mátyás, L. and M.N. Harris (1993), "A Comparative Analysis of Different Monte Carlo Methods", Working Paper 10/93, Department of Econometrics, Monash University.

Small, K.A. and C. Hsiao (1985), "Multinomial Logit Specification Tests", *International Economic Review*, **16**, 471-486.

Tse, Y.K. (1987), "A Diagnostic Test for the Multinomial Logit Model", *Journal of Business & Economic Statistics*, **5**, 283-286.

Zhang, J. and S.D. Hoffman (1993), "Discrete-Choice Models: Testing the IIA Property", *Sociological Methods & Research*, **22**, 193-213.

Table 1: Number of Observations per Outcome.

Outcome	n_j
Discharged	90
Probation	80
Community Service Order	102
Suspended Imprisonment	62
Immediate Imprisonment	95
Fined	260

Table 2: Values of Choice Set Partitioned Tests.

Outcome Dropped	HM	MTT	SH	H
Discharged	5.5678	0.3120	28.4113	57.1906
Probation	14.4877	0.2921	28.7131	57.3868
Community Service Order	2.9612	0.2074	29.4664	57.5110
Suspended Imprisonment	2.3113	0.4675	32.7032	63.8486
Immediate Imprisonment	50.2097	0.1555	28.9470	58.2089
Fined	-20.7715	2.9518	19.0690	37.8036

Note: All these tests have asymptotic $\chi^2(20)$ distributions with an (asymptotic) 5% critical value = 31.41.

Tests against DOGIT: $\lambda_{LM} = 0.2953 \overset{a}{\sim} \chi^2(6)$, $\lambda_{LMMP} = -0.1590 \overset{a}{\sim} N(0,1)$.

[5% asymptotic critical values 12.59 and 1.645 respectively]

Table 3: Empirical Size Performance for a Nominal Size of 5%.

Outcome Dropped	HM	MTT	SH	H
Discharged	0.036	0.000	0.084	0.763
Probation	0.059*	0.000	0.072	0.750
Community Service Order	0.053*	0.000	0.075	0.753
Suspended Imprisonment	0.041*	0.000	0.089	0.786
Immediate Imprisonment	0.030	0.000	0.084	0.758
Fined	0.047*	0.000	0.086	0.760

Tests against DOGIT: $\lambda_{LM} = 0.004$, $\lambda_{LMMP} = 0.000$.

* indicates that 95% confidence interval for estimated size contains nominal size.

Table 4: Empirical 5% Critical Values for the Tests.

Outcome Dropped	HM	MTT	SH	H
Discharged	23.4449	0.4926	33.9267	66.8484
Probation	35.1879	1.2866	32.9364	65.4635
Community Service Order	32.1753	1.4375	33.3754	66.5297
Suspended Imprisonment	25.4402	0.8313	34.4868	68.1514
Immediate Imprisonment	17.1097	0.5314	33.5588	66.4270
Fined	29.2051	2.7853	33.3408	64.0253

Tests against DOGIT: $\lambda_{LM} = 3.5886$, $\lambda_{LMMP} = 0.3497$.

APPENDIX: Descriptive Statistics and Estimation Results.

n = 689, corresponding to courts 2, 3 and 6 of the NACRO study.

SENTENCE: description in Table 1.

OSCORE: Offending Score Index - min. 3, max. 19, mean 11.083, s.d. 4.0639.

EMPT: 0/1 indicator of employment status (=1 if in employment) - 242 (0.3512) 1's.

CRT3: 0/1 indicator of court 3 (= 1 if in court 3) - 248 (0.3599) 1's.

CRT6: 0/1 indicator of court 6 (= 1 if in court 6) - 231 (0.3353) 1's.

Maximum Likelihood Estimates of Multinomial Logit Model:

Outcome	Coefficient	Estimate	Std. Error
Discharged	Constant	0.4364	0.4245
	OSCORE	-0.1048	0.0377
	EMPT	-0.9868	0.2744
	CRT3	-0.0023	0.3114
	CRT6	-0.8723	0.3889
Probation	Constant	-2.1263	0.5135
	OSCORE	0.2391	0.0429
	EMPT	-1.1664	0.3153
	CRT3	-1.5813	0.3442
	CRT6	-1.8100	0.3874
Community Service Order	Constant	-4.1414	0.6016
	OSCORE	0.3470	0.0457
	EMPT	-1.2486	0.3176
	CRT3	-1.0690	0.4107
	CRT6	0.2296	0.3754
Suspended Imprisonment	Constant	-7.5168	0.9887
	OSCORE	0.6552	0.0755
	EMPT	-0.4660	0.3612
	CRT3	-4.1694	0.6069
	CRT6	-1.8994	0.4082
Immediate Imprisonment	Constant	-9.5396	1.055
	OSCORE	0.8494	0.0790
	EMPT	-2.0181	0.4085
	CRT3	-4.1026	0.5139
	CRT6	-1.8315	0.3990

Note: Coefficient estimates normalized on Fined category.

Log-likelihood = -865.145, Restricted (slopes=0) Log-likelihood = -1141.211.

Frequencies of Actual and Predicted Outcomes:

Predicted → Actual	Disc.	Prob.	C.S.O.	S.I.	I.I.	Fined	Total
Discharged	8	5	3	1	2	71	90
Probation	1	12	5	4	11	47	80
C.S.O.	0	4	24	4	26	44	102
S.I.	0	5	11	11	22	13	62
I.I.	0	2	17	9	60	7	95
Fined	9	5	19	10	7	210	260
Total	18	33	79	29	128	392	689

Percentage of correct predictions = 47.17.

Note: Predicted outcome has maximum probability.

Percentage of Negative Hausman-McFadden Test Statistics in 1000 Replications:

Outcome Dropped	% negative
Discharged	42.4
Probation	25.0
Community Service Order	24.2
Suspended Imprisonment	32.6
Immediate Imprisonment	36.4
Fined	26.1