Vijverberg, Wim P.

Working Paper

Testing for IIA with the Hausman-McFadden test

IZA Discussion Papers, No. 5826

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Vijverberg, Wim P. (2011) : Testing for IIA with the Hausman-McFadden test, IZA Discussion Papers, No. 5826, Institute for the Study of Labor (IZA), Bonn, https://nbn-resolving.de/urn:nbn:de:101:1-201107133399

This Version is available at:
https://hdl.handle.net/10419/51633

DISCUSSION PAPER SERIES

IZA DP No. 5826

Testing for IIA with the Hausman-McFadden Test

Wim Vijverberg

June 2011

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Testing for IIA with the Hausman-McFadden Test

Wim Vijverberg
CUNY Graduate Center
and IZA

Discussion Paper No. 5826
June 2011

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5826
June 2011

ABSTRACT

Testing for IIA with the Hausman-McFadden Test*

The Independence of Irrelevant Alternatives assumption inherent in multinomial logit models is most frequently tested with a Hausman-McFadden test. As is confirmed by many findings in the literature, this test sometimes produces negative outcomes, in contradiction of its asymptotic χ^2 distribution. This problem is caused by the use of an improper variance matrix and may lead to an invalid statistical inference even when the test value is positive. With a correct specification of the variance, the sampling distribution for small samples is indeed close to a χ^2 distribution.

JEL Classification: C12, C35

Keywords: multinomial logit, IIA assumption, Hausman-McFadden test

Corresponding author:

Wim Vijverberg
CUNY Graduate Center
365 5th Avenue
New York, NY 10016-4309
USA
E-mail: wvijverberg@gc.cuny.edu

* I appreciate the comments and contributions of seminar participants at Hunter College, Rutgers University at Newark, West Virginia University, and Wichita State University and in particular an insightful suggestion by Stratford Douglas.

1 Introduction

Multinomial logit models are valid under the Independence of Irrelevant Alternatives (IIA) assumption that states that characteristics of one particular choice alternative do not impact the relative probabilities of choosing other alternatives. For example, if IIA is valid, how I choose between watching a movie or attending a football game is independent of whoever is giving a concert that day. Violation of the IIA assumption complicates the choice model. Therefore, much is gained when the IIA assumption is validated.

A number of tests of the IIA exist. One test was devised by Hausman and McFadden (1984) as a variation of the Hausman (1978) test. It relies on the insight that (*i*) under IIA, the parameters of the choice among a subset of alternatives may be estimated with a multinomial logit model on just this subset or on the full set, though the former is less efficient than the latter, and (*ii*) if IIA is not true, the parameter estimates of the full set are inconsistent, whereas those of the subset are consistent provided that the subset is properly selected. This test is implemented simply by two (multinomial) logit estimations and an evaluation of the difference in the parameter estimates. Other tests include one designed by Small and Hsiao (1985), which builds on McFadden, Train, and Tye (1977); another test proposed by Hausman and McFadden (1984) based on an estimation of the nested logit model; tests based on regression-based statistics by McFadden (1987) and Small (1994); a nonparametric test by Zheng (2008); and a test by Weesie (1999) that will be discussed below.

For the purpose of this paper, let us denote the Hausman-McFadden *testing strategy* as the HM test and the popular *implementation* of this test as the $\tilde{H}$ test, to be contrasted with the $\hat{H}$ test that will be defined later on. The $\tilde{H}$ test is by far the most frequently used test, undoubtedly in no small part due to its simplicity. To illustrate, over the period from 1984 to February 2010, there were 388 studies published in the scholarly literature that cited the

Hausman and McFadden (1984) paper, 276 of which applied the $\tilde{H}$ test for a total of 433 test results (Table 1).[1] In fact, the use of this test is accelerating: as Table 1 shows, the test was applied in as many studies in the last five years as in the first 20 years. Furthermore, the test was included in comparative simulation studies by Fry and Harris (1996, 1998), Cheng and Long (2007), and Zheng (2008); Small and Hsiao (1985) and Zhang and Hoffman (1993) offer a numerical comparison as well.

One inconvenient finding about the $\tilde{H}$ test is that, even though its asymptotic distribution is χ^2 and its outcomes should be therefore strictly positive, in applied situations the test statistic sometimes yields negative outcomes. That negative values may occur has long been known, starting with Hausman and McFadden (1984, p.1226) who suggest that a negative outcome of the $\tilde{H}$ test may still be taken as support for the Null hypothesis. The literature clearly subscribes to this view: as Table 1 indicates, most studies do not reject IIA when $\tilde{H}$ is negative, and the Monte Carlo study of Cheng and Long (2007) continues this practice. However, negative values occur frequently: in the literature described in Table 1, 16.2 percent of the reported test values are negative, and one may suspect that the test statistic may well be negative among some of the 32 percent of the cases where authors merely make mention of the test without reporting its actual value.

The frequent occurrence of these negative values has unappreciated consequences. In this paper, I contend that $\tilde{H}$ is a test statistic with poor properties that, compounded by errors in inference that are due to its common implementation, has put a significant body of empirical research at risk. However, a better version of the HM test is readily available.

Through simulations as well as analytically derived small-sample distributions, I show

[1] These statistics are drawn from the Web of Science exclusively. Apart from the 276 studies that cited Hausman and McFadden (1984) as a reference for the test that was implemented, 47 studies used its contribution as a building block in their own pursuits; 61 made reference without a clear indication how it added value; and 4 listed the paper in the bibliography section but made no reference to it in the study itself. By comparison, McFadden, Train, and Tye (1977) was cited 76 times for all purposes combined; Small and Hsiao (1985) was cited 73 times; McFadden (1987) was cited 36 times; Small (1994) was cited 13 times; and Zheng (2008) has not been cited yet.

that the asymptotic χ^2 distribution offers a poor approximation to the small-sample distribution of $\tilde{H}$ for samples with 1000, 7500 or even 100,000 observations. The reason is that the $\tilde{H}$ test inserts a variance estimator of the parameter difference that is conceptually inconsistent with the research question. This has two consequences. First, this (estimated) variance matrix may become indefinite, according to our simulations not just occasionally but actually very frequently even in very large samples. When this happens, large differences between estimates of full and restricted models may equally well result in large positive or large negative values of the $\tilde{H}$ test, or small values for that matter. The fact that the variance matrix may be indefinite also means that the test statistic must not be accepted as "valid" whenever the variance happens to be positive definite. The inserted variance is part and parcel of $\tilde{H}$; therefore, $\tilde{H}$ has a distribution different from χ^2. This leads us to the second consequence: $\tilde{H}$ is a pivotal test statistic only asymptotically. In small samples, its sampling distribution depends on the sample at hand in two ways: (i) it uses parameter estimates of both the full and the restricted model, and (ii) it is conditional on the sample's choice outcomes. With much effort, an unconditional distribution under IIA may be derived, but the (estimated) sampling distribution remains sample-dependent and, as simulations show, is not robust.

The conceptually correct formulation of the HM test will be referred to as the $\hat{H}$ test. Its small-sample distribution does not deviate greatly from its asymptotic χ^2 sampling distribution: $\hat{H}$ is remarkably robust against small-sample deviations from non-normality in the parameter estimates. This formulation was mentioned almost parenthetically by Hausman and McFadden (1984), but it has not been taken seriously in the literature and is rarely implemented in practice (Table 1).[2] Overall, $\hat{H}$ performs significantly better than $\tilde{H}$.

[2]Hausman and McFadden (1984, p.1226) argued on the basis of work with $\hat{H}$ (at a time where computing power limited Monte Carlo research relative to today's standards) that negative values of $\tilde{H}$ may be taken as support for the Null hypothesis. For some data structures, $\hat{H}$ actually proves to have little power, which could explain this recommendation.

HM tests must flag samples for which differences in restricted and unrestricted parameter estimates are large. Across many simulations where IIA was violated, for those samples that were flagged by $\hat{H}$ at a 5 percent significance level, the size-corrected probability that $\tilde{H}$ was significant was only about 11 percent in small ($n = 1000$) samples and 21 percent in larger ($n = 7500$) samples—and $\tilde{H}$ actually fell in its 5-percent *left* tail about 6 percent of the time. Most of the time, $\tilde{H}$ was insignificant. Combined with the fact that $\tilde{H}$ is typically judged by its asymptotic χ^2 distribution and thus not evaluated properly, this strongly suggests that many among the 276 studies summarized in Table 1 may have made erroneous inferences: rejections of IIA may have been invalid, and acceptances of IIA may have been incorrect.

As mentioned, the inserted variance estimate is the cause of the troubles with the $\tilde{H}$ test. Weesie (1999) developed a general sandwich variance estimator that may be applied to HM tests for IIA as well. It ensures that the estimated variance is positive definite, yielding a formulation of the HM test that will be referred to as $\hat{H}^s$.[3] As this paper shows, in small samples $\hat{H}^s$ is exceedingly conservative and, even after size correction, is less powerful than $\hat{H}$.

In the following, Section 2 discusses the differences between the two versions of the Hausman-McFadden test and derives their analytical small-sample distributions, subject to certain restrictions. Section 3 examines these restrictions and the $\tilde{H}$ and $\hat{H}$ tests in a Monte Carlo context. Section 4 compares the analytical small-sample densities with their simulated counterparts. Section 5 describes Weesie's $\hat{H}^s$ test and the Monte Carlo evidence about it. Section 6 applies the three versions of the HM test, and Section 7 concludes.

[3]However, the literature has largely ignored (or been ignorant of) this test: over the period 2000-2010, only four studies that perform tests for IIA apply the $\hat{H}^s$ test that is based on Weesie (1999).

2 The HM test

This section reexamines the common implementation of the Hausman-McFadden test for IIA. It is shown to rely on a shortcut that leads to a specification of a variance estimator that contradicts the principle of comparison between the two multinomial logit models that are being compared by the test. However, the specification of the test can be improved upon with a simple modification. This section also develops tools to examine small-sample properties of the test statistics.

2.1 Two specifications of the HM test

Let sample $S_r = \{1, ..., n\}$ consist of n individuals who each choose from J alternatives in the choice set $C = \{1, ..., J\}$. Across the different samples ($r = 1, 2, \ldots, R$), the choice set remains the same. The index r is suppressed in the notation until the appropriate time.

The utility derived from alternative j is given by

$$U_{ji} = X_i\beta_j + \epsilon_{ji},\ j \in C,\ i \in S_r \tag{1}$$

where $X_i\beta_j$ represents the systematic component of utility and ϵ_{ji} the stochastic component. Alternative j is chosen if $U_{ji} > \max(U_{ki}, k \neq j)$. Accordingly, define $y_{ji} = 1$ if j is chosen and $= 0$ if not. Under the Null hypothesis of IIA, ϵ_{ji} is distributed Gumbel. Then, the log-likelihood function may be written as

$$\ln L_C = \sum_{i=1}^{n}\sum_{j\in C} y_{ji}X_i\beta_j - \sum_{i=1}^{n}\ln\left(\sum_{k\in C} e^{X_i\beta_j}\right) \tag{2}$$

Let $\beta_J = 0$ for standardization, and define $\beta = \left(\beta_1', \cdots, \beta_{J-1}'\right)'$. Its true value is denoted as β^0. Let $\hat{\beta}$ maximize $\ln L_C$. Then, asymptotically, $\hat{\beta}$ is distributed under H_0 as $N(\beta^0, \Sigma_C^0)$

with

$$\Sigma_C^0 = \left(-E_0\left[h_C(\beta^0)\right]\right)^{-1} \equiv \Sigma_C(\beta^0) \tag{3}$$

where block j, k of the hessian is given by

$$\left[h_C(\beta^0)\right]_{jk} = \frac{\partial^2 \ln L_C(\beta^0)}{\partial\beta_j \partial\beta_k'} = -\sum_{i=1}^{n} p_{ji}^0 \left(I(j=k) - p_{ki}^0\right) X_i X_i' \tag{4}$$

in which $p_{ji}^0 = e^{X_i\beta_j^0} / \sum_{k\in C} e^{X_i\beta_k^0}$ and I is the familiar indicator function. Equation (4) does not contain any random term, and so the expectation of it in equation (3) is the same expression. In estimation, $\hat{\beta}$ is substituted for β^0. We label the variance estimator therefore as $\hat{\Sigma}_C = \Sigma_C(\hat{\beta})$.

For this same sample, let us test for IIA by removing some of the alternatives from the choice set. The remaining choice set is denoted as D, which must be specified *a priori*, and let $D^\dagger = C \setminus D$ be D's complement. Without loss of generality it is assumed that D contains alternative J and that it is once again the basis for standardization. Let y_{Di} be the indicator that individual i selects any $j \in D$: $y_{Di} = \sum_{j\in D} y_{ji}$. Define β_D as the vector that stacks β_j for $j \in \{D \setminus \{J\}\}$, and extract β_D^0 similarly from β^0. Then the log-likelihood function for this estimation problem is:

$$\ln L_D = \sum_{i=1}^{n} y_{Di} \left(\sum_{j\in D} y_{ji} X_i \beta_j - \ln\left(\sum_{k\in D} e^{X_i\beta_j} \right) \right) \tag{5}$$

Let $\breve{\beta}_D$ be the estimator that maximizes $\ln L_D$.

In the common formulation of this estimation problem, the summation over i is made only over the subsample $S_r(D)$ of members of S_r who select an alternative $j \in D$, and y_{Di}

vanishes as it always equals 1. Denote this formulation as $\ln \tilde{L}_{Dr}$:

$$\ln \tilde{L}_{Dr} = \sum_{i \in S_r(D)} \sum_{j \in D} y_{ji} X_i \beta_j - \sum_{i \in S_r(D)} \ln \left(\sum_{k \in D} e^{X_i \beta_j} \right) \tag{6}$$

In other words, this becomes a regular MNL model estimated over a smaller sample and a smaller choice set. Taken as such, this leads to the conclusion that the asymptotic variance of $\breve{\beta}_D$ is given by

$$\tilde{\Sigma}^0_{Dr} = \left(-E_0 \left[h^0_{Dr}(\beta^0_D) \right] \right)^{-1} \equiv \tilde{\Sigma}_{Dr}(\beta^0_D) \tag{7}$$

where for $j, k \in D$:

$$\left[h^0_{Dr}(\beta^0_D) \right]_{jk} = \frac{\partial^2 \ln \tilde{L}_D(\beta^0_D)}{\partial \beta_j \partial \beta'_k} = - \sum_{i \in S_r(D)} q^0_{ji} \left(I(j = k) - q^0_{ki} \right) X_i X'_i \tag{8}$$

with $q^0_{ji} = e^{X_i \beta^0_j} / \sum_{k \in D} e^{X_i \beta^0_k} = p^0_{ji} / p^0_{Di}$, where $p^0_{Di} = \sum_{k \in D} p^0_{ki}$. As before, in estimation, $\breve{\beta}_D$ is substituted for β^0_D, and we label the variance estimator therefore as $\breve{\tilde{\Sigma}}_{Dr} = \tilde{\Sigma}_{Dr}(\breve{\beta}_D)$.

However, the likelihood function given in equation (5) yields a different expression for the variance of $\breve{\beta}_D$:

$$\Sigma^0_D = \left(-E_0 \left[h^0_D(\beta^0_D) \right] \right)^{-1} \equiv \Sigma_D(\beta^0) \tag{9}$$

where the expected value of block j, k of the hessian equals

$$\begin{aligned} E_D \left[\left[h^0_D(\beta^0_D) \right]_{jk} \right] = E_0 \left[\frac{\partial^2 \ln L_D(\beta^0_D)}{\partial \beta_j \partial \beta'_k} \right] &= E \left[- \sum_{i=1}^{n} y_{Di} q^0_{ji} \left(I(j = k) - q^0_{ki} \right) X_i X'_i \right] \\ &= - \sum_{i=1}^{n} p^0_{Di} q^0_{ji} \left(I(j = k) - q^0_{ki} \right) X_i X'_i \\ &= - \sum_{i=1}^{n} p^0_{ji} \left(I(j = k) - q^0_{ki} \right) X_i X'_i \end{aligned} \tag{10}$$

Note that p^0_{ji}, even for $j \in D$, is a function of the entire β^0 vector. Thus, Σ^0_D depends on

β^0. Accordingly, to estimate Σ_D^0, we insert $\hat{\beta}$: $\hat{\Sigma}_D = \Sigma_D(\hat{\beta})$.

What is the difference between Σ_D^0 and $\tilde{\Sigma}_{Dr}^0$? The answer lies in the recognition of ϵ_{ji} for all $j = 1, \ldots, J$ and all $i = 1, \ldots, n$ as the source of randomness. $\tilde{\Sigma}_{Dr}^0$ describes the variation in $\breve{\beta}_D$ that results from confronting members of the subsample $S_r(D)$ with different realization of the world, by assigning them different draws of ϵ_{ji} for $j \in D$ and restricting their choice to choice set D only. The membership of this subsample does not change; only what they choose from D is changing. On the other hand, Σ_D^0 describes the variation in $\breve{\beta}_D$ that results from assigning different draws of ϵ_{ji} for all $j \in C$ to the entire sample S_r, not just for $j \in D$. Some members of $S_r(D)$ will no longer choose $j \in D$, and some of $S_r(D^\dagger)$ will now choose $j \in D$. Given that sample S_r rather than the subsample $S_r(D)$ is the basis of the test for IIA, this second characterization of the randomness in $\breve{\beta}_D$ is conceptually preferable to the first.

It could be argued that $\breve{\tilde{\Sigma}}_{Dr}$ and $\hat{\Sigma}_D$ are merely two alternative estimators of Σ_D^0. Indeed, they are, with two differences between them. First, they rely on different estimators of β, namely $\breve{\beta}_D$ and $\hat{\beta}$, respectively. Second, rewrite equation (8) as

$$\begin{aligned} \left[h_{Dr}^0\right]_{jk} &= -\sum_{i \in S_r(D)} q_{ji}^0 \left(I(j=k) - q_{ki}^0\right) X_i X_i' \\ &= -\sum_{i=1}^{n} y_{Di} q_{ji}^0 \left(I(j=k) - q_{ki}^0\right) X_i X_i' = \left[h_D^0\right]_{jk} \end{aligned} \tag{11}$$

This highlights the fact that the expectation in equation (7) is conditional on selection of D (i.e., on y_D) by members of S_r. Thus, it is more accurate to write $\tilde{\Sigma}_{Dr}^0 = \tilde{\Sigma}_D(\beta_D^0, y_{Dr})$. Then, compare the second line of equation (10) with equation (11): to estimate Σ_D^0, y_{Di} is substituted for p_{Di}. Asymptotically, this does not matter since $\breve{\tilde{\Sigma}}_D^0$ converges to Σ_D^0 as n goes to ∞ (Appendix A). But in a small sample y_{Di} is a poor substitute for p_{Di}.[4]

[4]Of course, if the objective is to study the choice of alternatives from choice set D and samples are drawn accordingly, the estimator $\tilde{\Sigma}_D(\breve{\beta}_D)$ is appropriate, just as $\Sigma_C(\hat{\beta})$ is for what is considered to be the "full" model in this paper.

This brings us to formulations of the HM test statistic. Let $\hat{\beta}_D$ be the estimator of β_D that is extracted from $\hat{\beta}$ of the full MNL model; its variance $\Sigma^0_{C,DD}$ consists of the corresponding submatrix of Σ^0_C. Define $\hat{\delta}_D = \breve{\beta}_D - \hat{\beta}_D$. If IIA applies, the mean of $\hat{\delta}_D$ equals 0, and its asymptotic variance equals

$$\Omega^0_D = \Sigma^0_D - \Sigma^0_{C,DD} \equiv \Omega(\beta^0) \tag{12}$$

In principle, the HM test statistic is then given by $H = \hat{\delta}'_D {\Omega^0_D}^{-1} \hat{\delta}_D$. H is actually infeasible since Ω^0_D is unknown, but it serves a purpose in the Monte Carlo study. In line with the previous discussion, Ω^0_D may be estimated with $\hat{\Omega}_D = \Omega(\hat{\beta})$ or with $\breve{\tilde{\Omega}}_{Dr} = \tilde{\Sigma}_D(\breve{\beta}_D, y_{Dr}) - \Sigma_{C,DD}(\hat{\beta})$. Ω^0_D and $\hat{\Omega}_D$ are positive definite (see Appendix B), but nothing can be said about $\breve{\tilde{\Omega}}_{Dr}$. On the basis of these, the test statistics are denoted as $\hat{H} = \hat{\delta}'_D \hat{\Omega}^{-1}_D \hat{\delta}_D$ and $\tilde{H} = \hat{\delta}'_D \breve{\tilde{\Omega}}^{-1}_{Dr} \hat{\delta}_D$. $\tilde{H}$ is the common formulation of the HM test that is used in nearly every study that tests for IIA. $\hat{H}$ uses a conceptually preferred estimator of Ω^0_D.

Hausman and McFadden (1984) mention $\hat{H}$ almost parenthetically and then only as a fix in case $\breve{\tilde{\Omega}}_{Dr}$ is indefinite and as a way to motivate that a negative outcome of $\tilde{H}$ may well indicate non-rejection of the Null hypothesis. The Monte Carlo analysis in Section 3 will shed light on the validity of this assessment. In particular, $\tilde{H}$ is negative so often that it would not be right to discard such outcomes and proceed to another test. Moreover, the differences between $\breve{\tilde{\Sigma}}_{Dr}$ and $\hat{\Sigma}_D$ have the potential to cause a significant distortion in $\tilde{H}$ relative to $\hat{H}$, which the literature has not yet recognized to be problematic.[5]

At a few points in the analysis below, we will compare $\hat{H}$ and $\tilde{H}$ with $H^e = \hat{\delta}'_D (\Omega^e_D)^{-1} \hat{\delta}_D$, where Ω^e_D is the empirical covariance matrix of simulated $\hat{\delta}_D$ values generated under the null hypothesis, i.e., a simulation-based estimator of Ω_D that is independent of any model

[5]Instead of replacing $\breve{\tilde{\Sigma}}_{Dr}$ with $\hat{\Sigma}_D$, we could replace $\hat{\Sigma}_C$ by a variance of $\hat{\beta}$ computed under conditional sampling that assigns draws of ϵ_{ji} for all $i = 1, \ldots, n$ such that the composition of subsamples $S_r(D)$ and $S_r(D^\dagger)$ stays the same. Under such conditional sampling, $\hat{\beta}$ would actually be biased and inconsistent (Appendix C). Thus a HM test statistic that is derived from a conditional sampling approach would be flawed from the start.

structure. Because it is computation-intensive, it is not at the center of our interest, but it does offer a yardstick to evaluate outliers of the HM statistics without the structure that accompanies $\hat{\Omega}_D$ or $\breve{\breve{\Omega}}_{Dr}$. Moreover, in Section 5, we will examine the consequences of estimating Ω_D^0 with the sandwich estimator proposed by Weesie (1999).

2.2 Small-sample distribution of the HM tests

As $\tilde{\Omega}_{Dr}$ is not guaranteed to be positive definite—and indeed frequently is indefinite—, the small-sample distribution may well deviate substantially from the asymptotic χ^2 distribution. Let us therefore develop tools to analyze the small-sample properties of the distribution of the HM tests.

It is beneficial to reorder β such that $\beta = \left(\beta'_{D^\dagger}, \beta'_D\right)'$, such that the parameters associated with D appear at the bottom; note that $\beta_J = 0$ because of standardization. Write $\hat{H}$ in terms of $\left(\hat{\beta}', \breve{\beta}'_D\right)'$:

$$\hat{H} = \begin{pmatrix} \hat{\beta} \\ \breve{\beta}_D \end{pmatrix}' R'_D \left(R_D \hat{V}_D R'_D\right)^{-1} R \begin{pmatrix} \hat{\beta} \\ \breve{\beta}_D \end{pmatrix} \tag{13}$$

where $\hat{V}_D$ is the estimated variance of $\left(\hat{\beta}', \breve{\beta}'_D\right)'$:

$$\hat{V}_D = \begin{pmatrix} & \hat{\Sigma}_C & & \hat{\Sigma}_{C,DD^\dagger} \\ & & & \hat{\Sigma}_{C,DD} \\ \hat{\Sigma}_{C,D^\dagger D} & \hat{\Sigma}_{C,DD} & & \hat{\Sigma}_D \end{pmatrix}. \tag{14}$$

and where $R = (0 \;\; I_D \;\; -I_D)$, such that I_D is an identity matrix conformable with β_D and 0 is a matrix of zeroes. Thus $\left(R_D \hat{V}_D R'_D\right) = \hat{\Omega}_D$. $\tilde{H}$ may be rewritten in the same way, using $\breve{\breve{V}}_{Dr}$ instead of $\hat{V}_D$ where $\breve{\breve{V}}_{Dr}$ is similar to equation (14) with $\breve{\breve{\Sigma}}_{Dr}$ substituted for $\hat{\Sigma}_D$.

Since $\breve{\breve{V}}_D \neq \hat{V}_D \approx \text{Var}\left(\left(\hat{\beta}', \breve{\beta}'_D\right)'\right)$, we may call on the procedure developed by Imhof (1961) to evaluate the distribution under the null hypothesis. A precise description of this

procedure is useful for an understanding of the small-sample behavior of the HM test statistics. In general terms, let an m-dimentional vector $\hat{\delta}$ be distributed $N(\delta, \Omega)$. Consider the quadratic form $H = \hat{\delta}' A \hat{\delta}$. Use the Cholesky decomposition to define P such that $\Omega^{-1} = PP'$ and $P'\Omega P = I$. Then, rewrite H:

$$H = \hat{\delta}' A \hat{\delta} = \hat{\delta}' P P^{-1} A (P')^{-1} P' \hat{\delta} = \hat{\delta}' P E \Theta E' P' \hat{\delta} = \hat{\eta}' \Theta \hat{\eta} = \sum_{j=1}^{m} \theta_j \hat{\eta}_j^2 = \sum_{j=1}^{m} \theta_j \chi^2(1, \mu_j^2) \quad (15)$$

where the third equality uses the singular value decomposition such that E and $\Theta = \mathrm{diag}(\theta_j)$ are the matrices of eigenvectors and eigenvalues of $P^{-1}A(P')^{-1}$, and where $\eta = C'P'\hat{\delta}$ is distributed $N(\mu, I_m)$ with $\mu = E'P'\delta$.[6] Provided that all eigenvalues are unique, H is distributed as a weighted sum of noncentral χ^2 random variables with degrees of freedom equal to 1 and noncentrality parameter μ_j^2.[7]

In the case of $\hat{H}$ under the null hypothesis of IIA, we have $\delta = 0$, $A = \hat{\Omega}_D$ and $\Omega = \Omega_D^0$. Thus, $\mu_j = 0$ for all j. Since Ω_D^0 is estimated by $\hat{\Omega}_D$, we have $P^{-1}A(P')^{-1} = I$, $\Theta = I$, and the estimated sampling distribution of $\hat{H}$ is χ^2. In the case of $\tilde{H}$, $\delta = 0$, $A = \breve{\tilde{\Omega}}_{Dr}$ and $\Omega = \Omega_D^0$. Since $\breve{\tilde{\Omega}}_{Dr} \neq \hat{\Omega}_D$, $\tilde{H}$ is distributed as a weighted sum of $\chi^2(1)$ variates. Moreover, whenever $\breve{\tilde{\Omega}}_{Dr}$ is an indefinite matrix, some of the eigenvalues are negative. In such a case, the estimated distribution of $\tilde{H}$ has a left tail stretching to $-\infty$. Unusually large values of $\hat{\delta}$ may therefore result in outcomes of $\tilde{H}$ in the right as well as the the left tail: the rejection range should be not merely the right tail. We return to this issue at the end of Section 3.2 when the quantitative importance of the left tail has become clearer.

Next, let us consider the sampling distribution under an alternative hypothesis. Many conditions could lead to a violation of IIA, but we employ the nested logit model as the

[6] By implication, Θ is also the matrix of eigenvalues of $A\Omega$.

[7] In case eigenvalues appear with multiplicity greater than 1, the associated $\hat{\eta}_j^2$ are combined into a single χ^2 with degrees of freedom equal to the multiplicity of that eigenvalue and a noncentrality parameter equal to the sum of the associated μ_j^2. In our analysis of $\tilde{H}$ (and of $\hat{H}$ under non-IIA), eigenvalues happen to be distinct.

model that is valid under the alternative hypothesis.[8] According to the nested logit model, the probability that choice j is selected equals

$$\begin{aligned} P(y_i = j|X_i) = p_{ji} &= \frac{e^{X_i\beta_j}}{F_i} \text{ for } j \in D^\dagger \\ &= \frac{e^{\lambda I_i}}{F_i}\frac{e^{X_i\beta_j/\lambda}}{E_i} \text{ for } j \in D \end{aligned} \tag{16}$$

where $I_i = \ln \sum_{j\in D} e^{X_i\beta_j/\lambda}$, $F_i = \sum_{j\in D^\dagger} e^{X_i\beta_j} + e^{\lambda I_i}$, and $E_i = e^{I_i}$.

The likelihood functions that are maximized are given by equations (2) and (5). Let the gradients be given by $g_C(\hat{\beta})$ and $g_D(\breve{\beta}_D)$, respectively. Approximate g_C at β^a, which is chosen such that $E[g_C(\beta^a)] = 0$ and thus equals the vector to which $\hat{\beta}$ converges as $n \to \infty$:

$$0 = g_C(\hat{\beta}) \approx g_C(\beta^a) + h_C(\beta^a)\left(\hat{\beta} - \beta^a\right) \tag{17}$$

Approximate g_D at β^b_D, which is chosen such that $E[g_D(\beta^b_D)] = 0$:

$$0 = g_D(\breve{\beta}_D) \approx g_D(\beta^b_D) + h_D(\beta^b_D)\left(\breve{\beta}_D - \beta^a_D\right). \tag{18}$$

In view of equation (16), we have $\beta^b_D = \beta^0_D/\lambda^0$ if the test specifies D correctly. The following results are straightforwardly derived:

$$E[g_C(\beta^a)g_C(\beta^a)'] = -h^0_C \tag{19}$$

$$E[g_D(\beta^b_D)g_D(\beta^b_D)'] = -h_D(\beta^b_D) \tag{20}$$

[8]For example, violations of IIA may be related to arbitrary correlations among the ϵ_j, to nonlinearity among linearly-included explanatory variables or to omitted explanatory variables; see Train (2009) for a good discussion. The nested logit model asserts that the only form of misspecification in the MNL model is the failure to account for a specific type of correlation among the ϵ_j.

$$E[g_C(\beta^a)g_D(\beta_D^b)'] = \begin{pmatrix} 0 \\ -h_D(\beta_D^b) \end{pmatrix} \tag{21}$$

where h_C^0 is the same as the right hand side of equation (4) with probabilities p_{ji}^0 evaluated through equation (16) at (β^0, λ^0). Thus,

$$V_D = \begin{pmatrix} h_C(\beta^a)^{-1} & 0 \\ 0 & h_D(\beta_D^b)^{-1} \end{pmatrix} \begin{pmatrix} -h_C^0 & 0 \\ 0 & -h_D(\beta_D^b) \\ -h_D(\beta_D^b) & -h_D(\beta_D^b) \end{pmatrix} \begin{pmatrix} h_C(\beta^a)^{-1} & 0 \\ 0 & h_D(\beta_D^b)^{-1} \end{pmatrix} \tag{22}$$

Since $h_C^0 \neq h_C(\beta^a)$, V_D no longer simplifies neatly, and in its estimated form, $R\hat{V}_D R'$ is no longer equal to $\hat{\Omega}_D$. The sampling distribution of both $\hat{H}$ and $\tilde{H}$ are weighted sums of non-central $\chi^2(1, \mu_j^2)$ distributions, with μ derived from $\delta = \beta_D^b - \beta_D^a$.

These tools are predicated on the assumption that $\hat{\beta}$ and $\breve{\beta}_D$ are normally distributed with means equal to β^a and β_D^b respectively and variances that are stated above. In the following section, we first report on Monte Carlo simulations that put a face on the analytical results and in the process examine the validity of this normality assumption in an ideal Monte Carlo world, and then evaluate the analytical sampling densities under the null and alternative hypotheses.

3 Monte Carlo evidence

3.1 Design

To examine the performance of the different versions of the Hausman-McFadden statistic, we employ three sets of data for both $J = 3$ and $J = 4$ in small and large sample versions. The first dataset is synthetic and is typical among Monte Carlo studies: it contains 1000 observations in the small sample version and 7500 observations in the large sample version;

the two explanatory variables are draws of standard normal variables with a correlation of about 0.48; and values for β are chosen such that one alternative is relatively dominant (Table 2). The second dataset is synthetic as well and is inspired by Cheng and Long (2007): with either 1000 or 7500 observations, its first variable is a uniform draw; its second is a normal draw added to the first variable; its third is a $\chi^2(1)$ draw added to the first variable; and the weights in this construction are chosen such that the correlation between the first and second variable is high (0.75) and moderate between the third and the first two (about 0.30). Thus, the second dataset is characterized by both multicollinearity and skewness. Again, values for β are chosen such that one alternative is relatively dominant. The third dataset derives from an analysis of employment activities among men in Côte d'Ivoire in urban areas other than Abidjan (Vijverberg, 1993). The explanatory variables consist of years of schooling and of apprenticeship, age, age squared, and four dummy variables (marital status, citizenship, and two time dummies). For β, we use parameters that are estimated from a MNL model of the actual employment choice in this sample: for $J = 3$ we include wage employment, non-farm self-employment, and farming, with $N = 1118$; for $J = 4$ we add non-employed men, such that $N = 1480$. A large sample version of this dataset stacks copies of this small sample until N equals approximately 7500: $N = 7826$ for $J = 3$, and $N = 7400$ for $J = 4$.

To examine behavior under the Null hypothesis of IIA, we generate $R = 5000$ runs; to study power, we use $R = 2000$. Across all runs, values of the explanatory variables are the same. As mentioned, the implementation of the Hausman-McFadden test is preceded by the selection of the restricted choice set D: the alternatives in D are indicated in the tables as well as by means of subscripts to H in the discussion.

3.2 Size

Let us start out with H, the version of the HM test that is formulated with the theoretical variance Ω_D^0. If the small-sample distribution of $\hat{\delta}_D$ is close to the asymptotic normal distri-

bution with mean 0 (as implied by IIA) and variance Ω_D^0, H should follow a distribution that is close to its asymptotic χ^2 distribution. Table 3 examines this by means of empirical sizes at nominal sizes of 10, 5 and 1 percent, and by a goodness-of-fit test.[9] Panel A describes the case of $J = 3$ for each of the three possible specifications of D, denoted as "12" which indicates inclusion of alternatives 1 and 2, "13" and "23". Panel B illustrates the results for $J = 4$ with a limited but representative selection of choice sets. It turns out that the sampling distribution of H deviates greatly from its asymptotic one for $N = 1000$, and for $N = 7500$ differences may still be large as well. How can this be, since H uses the correct theoretical (but unknown) variance? Tests show that, individually, most elements of $\hat{\beta}$ and $\breve{\beta}_D$ appear to be normally distributed,[10] but more than a few are significantly biased. Joint tests for normality reported in Appendix D indicate that, for $D = \{12\}$ and $D = \{123\}$ for each Set and especially in small samples, the estimates of $\hat{\beta}$ and $\breve{\beta}_D$ are biased, usually have a variance that deviates from the theoretical variance, and also differ in skewness and kurtosis from multivariate normality. All this applies *a fortiori* to $\hat{\delta}_D$ as well. Thus, (i) asymptotic p-values may indeed be inaccurate in small samples, and (ii) the accuracy of the small-sample tools developed in Section 2.2 may be in question since they are predicated on normality. Serious nonnormality may adversely impact the properties of $\tilde{H}$ and $\hat{H}$ and hamper the comparison between them. Fortunately, the robustness of $\hat{H}$ to nonnormality demonstrated in the simulations below should alleviate these concerns and give validity to the analytical investigation of properties of the test statistics in Section 4.

Table 4 describes the behavior of $\tilde{H}$, the common version of the HM test, under the Null hypothesis of IIA. In addition to the empirical sizes and the goodness-of-fit test, it reports the prevalence of indefinite $\breve{\breve{\Omega}}_{Dr}$-matrices, further distinguished by the likelihood of positive

[9]The goodness-of-fit test divides the χ^2 distribution into 20 equiprobable intervals and thus has a $\chi^2(19)$ distribution.

[10]Characteristics of the explanatory variables certainly matter. For example, estimated slopes that are related to the third ($\chi^2(1)$-related) variable of Set 2 are distinctly skewed.

or negative outcomes. Consider the first line of Table 4, the case of a small-sample Set 1 with $J = 3$ and $D = \{12\}$: $\tilde{H}$ suffers from large size distortions; for example, at a nominal size of 10 percent, the actual size equals 4.5 percent. Fully 40.8 percent of the $\tilde{H}$ values are negative, and $\tilde{\Omega}$ fails to be positive definite in 76 percent of the runs. The goodness of fit test value of 3918 far exceeds the 1% critical value of 36.19 and formally invalidates the use of the asymptotic χ^2 distribution to evaluate $\tilde{H}$.[11]

Results are similar for every simulation of $\tilde{H}$. In all three datasets, negative values of $\tilde{H}$ occur frequently, and in the vast majority of replications $\breve{\tilde{\Omega}}_{Dr}$ is indefinite. In fact, for $J = 3$ at least one of the diagonal elements of $\breve{\tilde{\Omega}}_{Dr}$ is negative for nearly 70 percent of the simulated values. The number of alternatives in the choice set matters little: the results for $J = 4$ are similar. Moreover, when the sample size grows to $N = 7500$, these adverse features do not improve much. Size distortions remain often serious and unpredictable—sometimes the empirical size is greater, sometimes it is less.

Relative to H, the $\hat{H}$ test inserts an estimate $\hat{\Omega}^0_D$ for Ω^0_D, which constitutes a distortion that should cause a deviation between the sampling distributions of H and $\hat{H}$. Given what was observed about H, Table 5 yields a surprise: the distortion moves the sampling distribution leftward such that it is fairly characterized by a χ^2 distribution.[12] The empirical size corresponds well with the nominal size, and the goodness-of-fit statistics indicate much smaller deviations (if any) from the χ^2 distribution than the ideal H test.[13] There is no clear

[11] In regard to the goodness-of-fit test, negative values of $\tilde{H}$ are accumulated in the first interval, although in principle they already are refuting the validity of applying the asymptotic χ^2 distribution.

[12] For $J = 3$, there are three configurations of D; for $J = 4$ there are 10. For six of the 39 small sample tests and three of the 39 large sample ones, all with Sets 1 or 2, we encountered a few negative outcomes of $\hat{H}$ that in each case amounted to less than 1 percent of the runs, and in a few other cases $\hat{\Omega}_D$ was found to be indefinite even when $\hat{H}$ came out positive (usually in only a few runs but as many as 12.4 percent for $D = \{134\}$ for Set 1 in a small sample setting). Mathematically, this is impossible since $\hat{\Omega}_D$ is a positive definite matrix. Numerical errors are to blame for this as $\hat{\Omega}_D$ is nearly singular: detailed analysis of these cases indicates that minor variations (in the seventh decimal) of $\hat{\beta}$ and $\breve{\beta}_D$ can change $\hat{H}$ drastically and without meaning. See Appendix E for a further illustration.

[13] For large samples across the three Sets, only two of the 39 goodness-of-fit statistics are significant at the 1-percent level.

theoretical reason to expect this result. For one thing, non-normality of $\hat{\delta}_D$ (Appendix D) ought to be reflected in a deviation from χ^2. Furthermore, $\hat{\Omega}_D$ varies substantially between draws: $\hat{H}_D$ is not just a simple quadratic form in $\hat{\delta}_D$.[14] On the other hand, on average, $\hat{\Omega}_D$ appears to be closer to the variance of the draws of $\hat{\delta}_D$ than Ω_D^0.[15]

Across Tables 3, 4 and 5, results appear to be robust to variations in the structure of the data. Results differ between the three datasets only in minor ways: Set 2 tends to create larger deviations from the χ^2 distribution, and increasing the number of explanatory variables in Set 3 does not appear to harm–and may indeed help–the performance of the statistics. Thus, we may conclude with confidence that these results may be extrapolated to any type of data structure. Moreover, these tables indicate that deviations from the asymptotic χ^2 are larger when D includes only less frequently selected alternatives and thus $S_r(D)$ subsamples are smaller.

Values of $\hat{H}$ may be used to evaluate "outliers" of $\hat{\delta}_D$ and their associated values of the $\tilde{H}$ statistic. Traditionally, only large positive values of $\tilde{H}$ are considered extreme. Now, let us consider $\tilde{H}_{123}$ for Set 3 (small sample) in comparison with $\hat{H}_{123}$: of the draws that $\hat{H}_{123}$ identifies as outliers (in its 5-percent upper tail), only 5.6 percent are found in the 5-percent right tail of $\tilde{H}_{123}$. In fact, 50.8 percent of these outliers are associated with a negative outcome of $\tilde{H}_{123}$, and 6.4 percent fall in the 5-percent left tail of the distribution of $\tilde{H}_{123}$.[16] Thus, the rejection range of the empirical distribution of $\tilde{H}$ is in doubt: because of

[14]The coefficient of variation of the diagonal elements equals about 0.5 for slope coefficients and 1.1 for the intercept, and off-diagonal elements vary similarly.

[15]For example, consider H_{123}^e for a small-sample Set 3 with $J = 4$ as an unstructured yardstick. The range from the first to the 99th percentile is $[7.1, 53.4]$ for H_{123}, much wider than $[6.0, 41.7]$ for H_{123}^e and $[7.2, 36.2]$ for $\hat{H}_{123}$. On the other hand, the correlation between H_{123} and H_{123}^e is 0.97, whereas the correlation between $\hat{H}_{123}$ and H_{123}^e is lower at 0.83, and the goodness of fit test of the distribution of H_{123}^e is 325, much worse than 51 for H_{123}.

[16]In contrast, 43.6 percent of these outlier draws of $\hat{H}_{123}$ yield a value of H_{123} in the 5-percent upper tail. In this light, also consider H_{123}^e: in the upper 5-percent tail area, there is an 81.6 percent overlap between H_{123} and H_{123}^e, 50.4 percent with $\hat{H}_{123}$, and only 3.6 percent with $\tilde{H}_{123}$. However, 6.4 percent of the draws in the upper 5-percent tail of H_{123}^e fall in the lower 5-percent tail of $\tilde{H}_{123}$.

the indefinite $\breve{\tilde{\Omega}}_{Dr}$, outliers of $\hat{\delta}_D$ may generate any value of $\tilde{H}_{123}$.[17]

3.3 Power

To examine power, we assume that the data generating process follows a nested logit structure. The nesting structure is parameterized with a parameter λ (e.g., see Hausman and McFadden (1984, Sec.2)): λ equals 1 for MNL, and a value of $\lambda \in (0, 1)$ implies positive correlation between alternatives, since the correlation equals $1 - \lambda^2$. We specify a value of λ such that violations of IIA should not be too hard to detect: $\lambda = 0.5$ and thus a correlation of 0.75. As for notation, a nesting structure 1(234) would indicate that alternatives 2, 3 and 4 are correlated and 1 stands independent.

Table 6 evaluates the power properties of $\tilde{H}$, the common implementation of the HM test, where power is computed at the empirical size of 5 percent, judged only by the right tail of the (empirical) distribution of the test and ignoring the left tail even if it covers a range of negative values. The number of alternatives equals $J = 3$ in Panel A and $J = 4$ in Panel B. For $J = 3$, all three nesting structures are considered: since some branches are more popular choices than others (Table 2), this varies the size of the subsample for which the restricted choice set is examined. For $J = 4$, we consider three structures, one with a correlation among two branches and two with a correlation among three branches.

Consider the first block of results, namely for Set 1, correlation structure (12)3, and tests based on $D = \{12\}$, $\{13\}$ and $\{23\}$. When $D = \{12\}$, the correlation structure is identified correctly but correlation is detected only 7.4 percent of the time. There is a 19.5 percent likelihood that $\tilde{H}$ comes out negative and 36.9 percent chance that $\breve{\tilde{\Omega}}_{Dr}$ is indefinite, in which case one is tempted to call the test inconclusive and thus perhaps not reject the Null.

[17] The $\tilde{H}$ test might be reformulated as a two-tailed test such that large negative values count as outliers as well: define the rejection range as $(-\infty, \tilde{H}_L] \cup [\tilde{H}_U, \infty)$ where $\tilde{H}_L$ marks the 5th percentile of the negative tail such that $F(\tilde{H}_L)/F(0) = 0.05$ with F being the cdf of $\tilde{H}$, and where $\tilde{H}_U$ denotes the 95th percentile of the positive tail such that $(1 - F(\tilde{H}_U))/(1 - F(0)) = 0.05$. However, the evidence about power in the next sections makes this a moot point anyway.

Power is actually higher when the correlation structure is incorrectly specified as $D = \{23\}$ but problems with a negative $\tilde{H}$ and an indefinite $\breve{\tilde{\Omega}}_{Dr}$ are more pervasive, which might strengthen the erroneous conclusion that correlation is absent.

All this is not peculiar to just the first set of results. In general, Table 6 suggests the following: (*i*) Even for this high degree of correlation, power is low even in large samples. (*ii*) Power is sometimes but not always higher if the nesting structure is identified correctly. (*iii*) Unless the nesting structure is identified correctly, $\tilde{H}$ is usually based on an indefinite $\breve{\tilde{\Omega}}_{Dr}$, especially when the number of alternatives grows, and all too frequently yields a negative test value, even when samples are large. (*iv*) Conforming to Hausman and McFadden (1984, p.1226), if indeed the test examines a restricted choice set that corresponds with the actual nesting structure, problems with $\breve{\tilde{\Omega}}_{Dr}$ and $\tilde{H}$ occur less frequently. Yet, it is not unusual for $\breve{\tilde{\Omega}}_{Dr}$ to be indefinite and for $\tilde{H}$ to turn out negative, apparently depending on the type of data and the size of the subsample that chooses from the restricted subset.

How does this compare with the power properties of $\hat{H}$? Consider Table 7. For small samples of Sets 1 and 2, power is not high either: across all alternatives with $J = 3$ or $J = 4$, the highest value is 0.296. In large samples, power ranges from 0.357 to 0.960 for correctly identified three-branch structures, but two-branch structures are still difficult to detect (power ranges from 0.081 to 0.679). The situation is better for Set 3 when D is selected correctly: when samples are small, power for $J = 3$ is 0.510, 0.709 and 0.592, respectively, and for $J = 4$ power equals 0.595, 0.816 and 0.763 respectively. For $N = 7500$, power exceeds 0.99. Thus, $\hat{H}$ performs significantly better than $\tilde{H}$.

$\hat{H}$ flags different simulation samples for violation of IIA than $\tilde{H}$. Across all small-sample Sets and branching structures, among the runs that produced a statistically significant value of $\hat{H}$, the probability that $\tilde{H}$ was statistically significant was only 12.7 percent for $J = 3$ and 8.9 percent for $J = 4$. For large samples, these percentages were still low at 23.7 and 18.9, respectively. Yet, across these four categories, the probability that the $\tilde{H}$ fell in its 5-percent

left tail was 4.4, 5.7, 6.9 and 5.6 percent, respectively. If these percentages would have been higher, this would suggest that a test for IIA through $\tilde{H}$ is truly a two-tailed test. Instead, it is hard to escape the conclusion that the information in $\tilde{H}$ is confounded by a large amount of statistical noise that leaves many large $\hat{\delta}_D$ unflagged.[18]

4 Analytical small-sample densities

A Monte Carlo evaluation of the sampling distribution of $\tilde{H}$ and $\hat{H}$ is computation-intensive. The Imhof procedure offers a faster analytical approach that, strictly taken, is valid only when the underlying assumptions are valid. As noted in Section 3.2, parameter estimates do exhibit non-normality in small sample contexts, but $\hat{H}$ appears to be quite robust to underlying non-normality. Therefore, we are now justified to use analytical tools in order to further clarify the difference between $\tilde{H}$ and $\hat{H}$.

As indicated by the subscript r in the notation of $\breve{\tilde{\Omega}}_{Dr}$, the estimated sampling distribution of $\tilde{H}$ under the null hypothesis is sample-dependent. This dependence is caused by the insertion of y_{Dr} and two different parameter estimates, namely $\hat{\beta}$ and $\breve{\beta}_D$, into $\breve{\tilde{\Omega}}_{Dr}$, causing sample-dependent variation in the eigenvalues θ_{jr}. The mean of the estimated sampling distribution under IIA equals $\sum_j \theta_{jr}$; e.g., for the simulated Set 3 ($J = 4$, small sample) and $D = \{123\}$, this mean averaged 7.16 with a standard deviation of 507.69 across 5000 replications. Let us illustrate this sample dependence graphically for this data structure: Figure 1 depicts the estimated Null distribution for the replication samples that generated means at the 10th, 50th and 90th percentile (-18.58, 7.83 and 35.37) and are referred to as the m10, m50 and m90 samples, respectively. These gray dashed, solid and dotted curves each deviate substantially from the black dashed curve that denotes the asymptotic $\chi^2(18)$ distribution, which itself has a mean of 18.

[18] A comparison with H^e further confirms this conclusion: for example, for a small-sample Set 3 with $J = 4$, H^e_{123} is essentially uncorrelated with $\tilde{H}_{123}$ (0.01) but positively correlated with $\hat{H}_{123}$ (0.68).

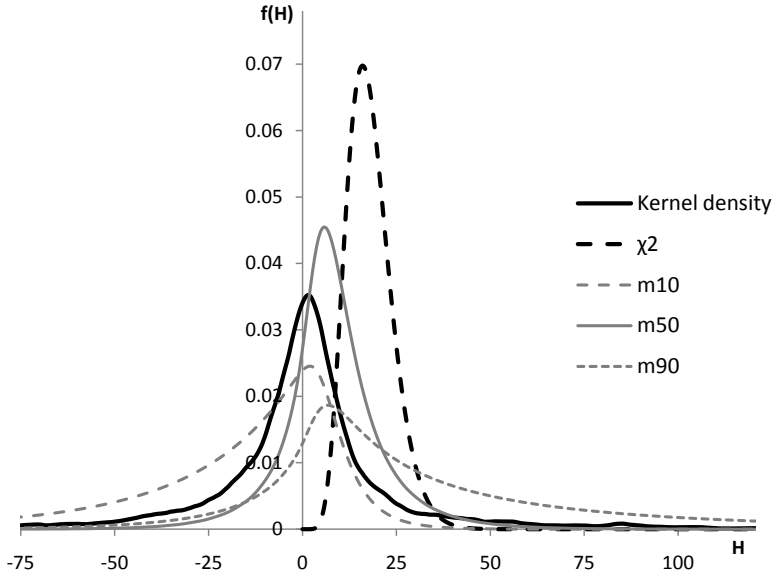


Figure 1: Estimated density functions of $\tilde{H}$ under the Null hypothesis

Note: Curves labeled m10, m50 and m90 are estimated densities on the basis of replication samples for which under the Null hypothesis the mean of the sampling distribution is at the 10th, 50th and 90th percentile, respectively. These density curves are computed for Set 3, small sample, $J = 4$ and $D = \{123\}$.

Ultimately, none of these sample-dependent densities can reliably identify the density of $\tilde{H}$ that indicates the spread of outcomes across many replications. Recall that $\breve{\breve{\Omega}}_{Dr} = \breve{\breve{\Sigma}}_{Dr} - \hat{\Sigma}_{C,DD}$: as argued in Section 2.1, the formulation of $\breve{\breve{\Sigma}}_{Dr}$ presumes that membership of subsample $S_r(D)$ is preserved across replications. Even if the formulation of $\hat{\Sigma}_{C,DD}$ is free of such restriction, this still implies that the unconditional density of $\tilde{H}$ is a mixture of these sample-dependent densities with weights proportional to the probability that y_{Dr} occurs as is observed.[19] In practice, a researcher has access to only one database and, without further Monte Carlo simulation (as was done here with Set 3), is unable to derive the unconditional

[19]Even this does not resolve sample dependence entirely since $\breve{\breve{\Omega}}_{Dr}$ depends on both $\hat{\beta}$ and $\breve{\beta}_D$.

density of $\tilde{H}$.

An estimate of this unconditional density is given by the kernel density in Figure 1. Only by a fortunate coincidence would a single sample-dependent density resemble the kernel density. The tails of the kernel density are actually exceedingly long: the first and 99th percentiles are at -314.5 and 334.1, respectively.

How does a transition towards a nested logit structure shift these sampling distributions? For each replication sample, we generate the same sets of uniform random numbers as under the multinomial model, transform them into correlated Gumbel variates, and determine the optimal choice y_j. β^0 and λ^0 are estimated with maximum likelihood on a nested logit model, and β^a is computed such that $E[\hat{g}_C(\beta^a)] = 0$ where $\hat{g}_C$ is a function of these estimates of β^0 and λ^0. Figure 2 illustrates the shift in the estimated sampling distributions as λ falls from 1 to 0.5, with correlation rising from 0 to 0.44, 0.64 and 0.75. Because the test statistic has so many moving parts ($\hat{\beta}$, $\breve{\beta}_D$ and especially y_{Dr}), the pace and direction of the shift is uneven. For example, for the m10 sample the left tail thins out quickly, but the right tail is thickest for $\lambda = 0.75$. For the m50 sample, the sampling distribution is unchanged for $\lambda = 75$, moves left instead of right for $\lambda = 0.60$ and develops a long left tail for $\lambda = 0.50$. For m90, the right tail empties out quickly, the distribution moves rightward when λ rises from 0.75 to 0.60 and then flattens for $\lambda = 0.50$. None of these gives a clear indication of rising power as λ decreases. Panel d sheds more clarity on this issue through kernel densities of 5000 simulated values of $\tilde{H}_{123}$. The sampling distribution does move steadily rightward, but the right tail is not becoming thicker. The empirical 5-percent critical value equals 58.63, and for $\lambda = 0.75$, 0.60 and 0.50, the tail probability equals only 0.055, 0.045 and 0.046.

The sampling distribution of $\hat{H}$ moves rightward in a more predictable way; see Figure 3. The estimated sampling distribution becomes sample-dependent because $\hat{\Omega}_D$ is no longer the same as the estimated variance of $\hat{\delta}_D$; compare equations (14) and (22). Illustrations with the m10, m50 and m90 replication samples show that the difference between them is not

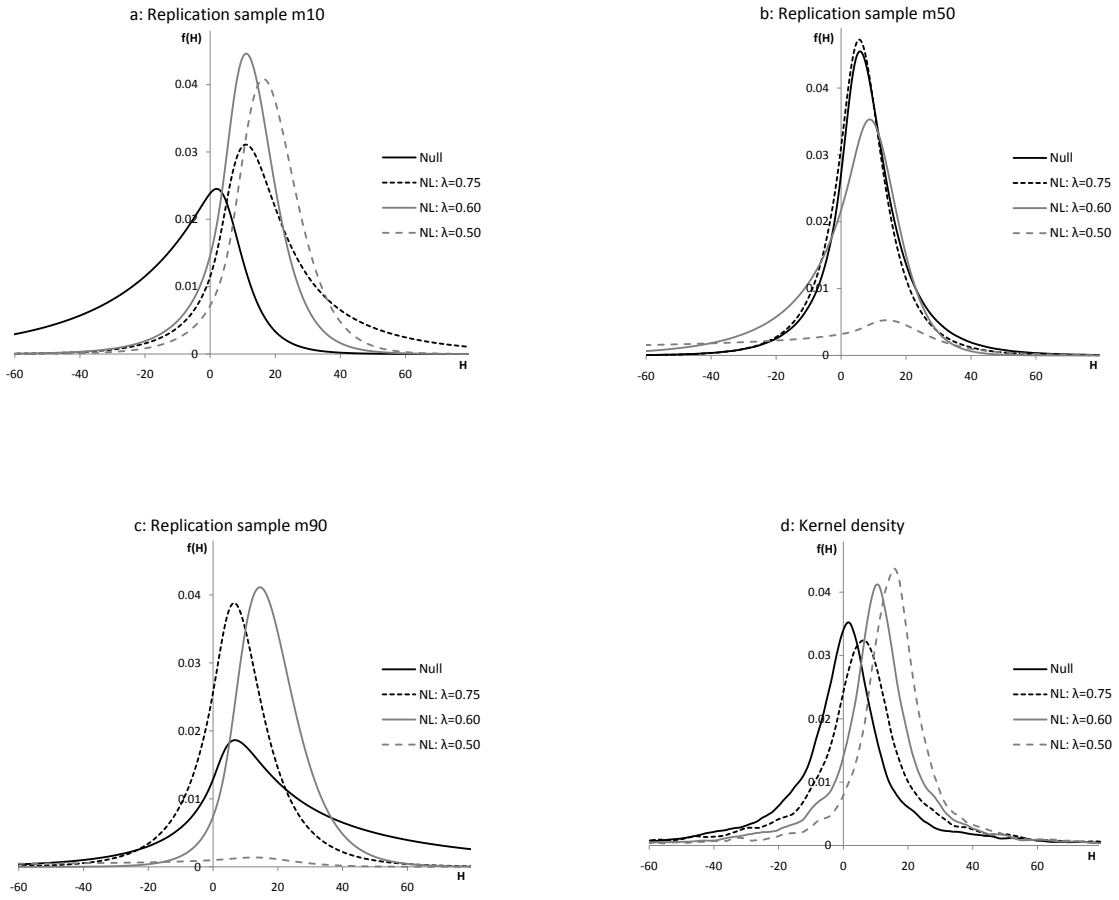


Figure 2: Density functions of $\tilde{H}$ under the null and alternative hypotheses

trivial, but in view of the critical value of 30.39, power clearly rises with falling λ for every replication sample. The kernel density illustrates the steady shift rightward as λ decreases. Note also that the density under the null hypothesis is not much different from the χ^2 density (dotted gray curve), even though the goodness of fit test (Table 5) pointed out a significant difference.

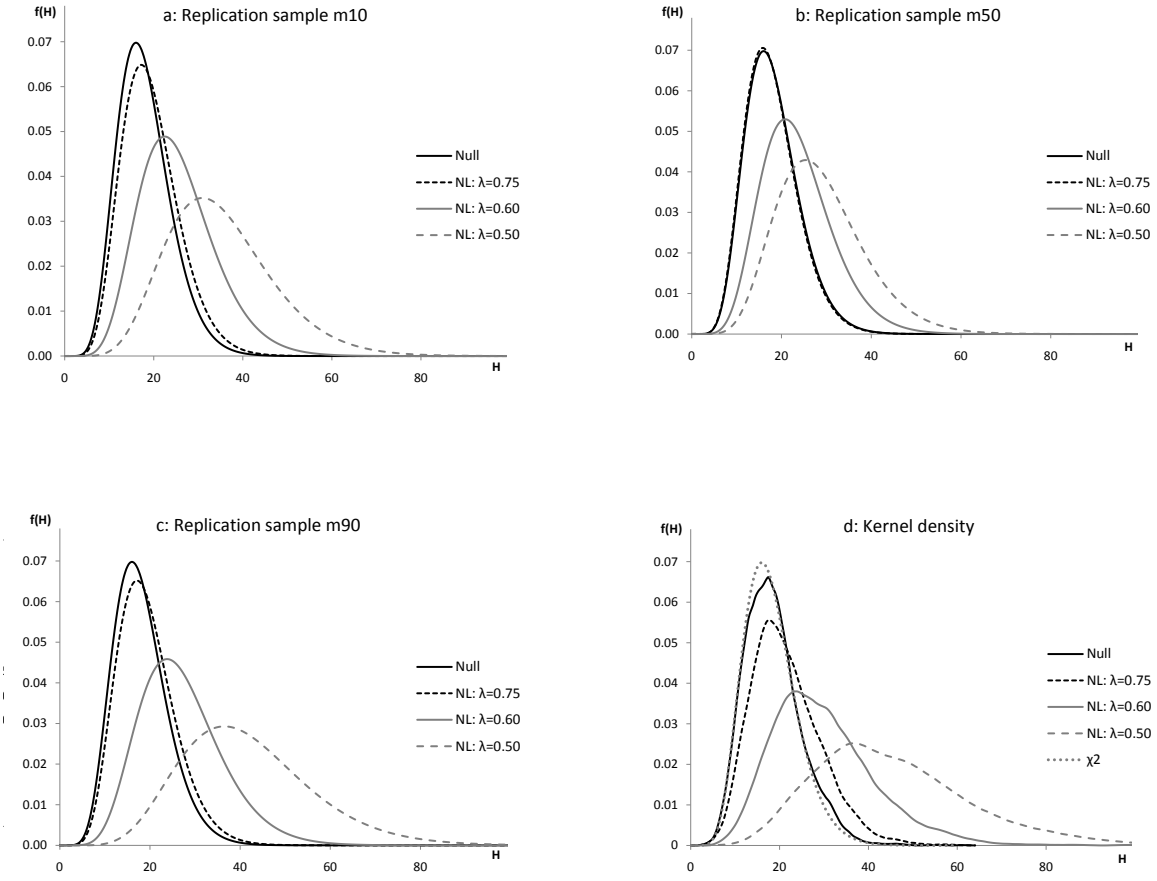


Figure 3: Density functions of $\hat{H}$ under the null and alternative hypotheses

5 A sandwich version of the HM test

The $\hat{H}^s$ test of Weesie (1999) replaces $\hat{V}_D$ in equation (14) with a sandwich estimator, as an application of a technique that applies more generally to tests on different sets of parameters that are drawn from the same sample. Thus, use Taylor expansions as in equations (17) and (18), and write the parameter vectors as

$$\begin{pmatrix} \hat{\beta} - \beta^0 \\ \tilde{\beta}_D - \beta_D^0 \end{pmatrix} = - \begin{pmatrix} h_C(\beta^0)^{-1} & 0 \\ 0 & h_D(\beta_D^0)^{-1} \end{pmatrix} \begin{pmatrix} g_C(\beta^0) \\ g_C(\beta_D^0) \end{pmatrix} \tag{23}$$

Let g_{Ci} and g_{Di} be the gradients of the respective log-likelihood functions for observation i. Then, the sandwich estimator of V_D is given by

$$\hat{V}_D^s = \begin{pmatrix} h_C(\hat{\beta})^{-1} & 0 \\ 0 & h_D(\breve{\beta}_D)^{-1} \end{pmatrix} \sum_{i=1}^{n} \begin{pmatrix} g_{Ci}(\hat{\beta}) \\ g_{Di}(\breve{\beta}_D) \end{pmatrix} \begin{pmatrix} g_{Ci}(\hat{\beta}) \\ g_{Di}(\breve{\beta}_D) \end{pmatrix}' \begin{pmatrix} h_C(\hat{\beta})^{-1} & 0 \\ 0 & h_D(\breve{\beta}_D)^{-1} \end{pmatrix} \tag{24}$$

Asymptotically, the middle term simplifies to an expression similar to that in equation (22) evaluated at β^0 and β_D^0, which underlies $\hat{V}_D$ in equation (14). In small samples, the sandwich estimator $\hat{V}_D^s$ may deviate from $\hat{V}_D$, which will cause the distribution of $\hat{H}^s$ to deviate from that of $\hat{H}$. Monte Carlo analysis must give insight into how large this deviation may be.

Table 8 shows size of $\hat{H}^s$ and must be compared with Table 5 for $\hat{H}$. In small samples, $\hat{H}^s$ suffers from serious size distortions: for example, the actual size at a nominal size of 5 percent is frequently below 1 percent. Even for large samples, the actual size is no larger than 3.8 percent. Thus, the critical value given by the χ^2 distribution is too high. The large goodness of fit statistics highlight the gap between the distribution of $\hat{H}^s$ and the asymptotic χ^2 distribution.

This size distortion leads to substantial underrejection of the Null hypothesis. For $J = 4$ with a small-sample Set 3, in the cases with $D = \{12\}$, $D = \{123\}$ and $D = \{234\}$ when the nesting structure with $\lambda = 0.50$ is correctly specified, we find nominal power equal to 0.063, 0.065 and 0.043, respectively, whereas actual power is 0.176, 0.247 and 0.240 (Table 9, lower left block). However, these are substantially below the power of $\hat{H}$ in same three cells of Table 7, equal to 0.595, 0.816, and 0.763, respectively. With a few exceptions, the power of $\hat{H}^s$ is lower than that of $\hat{H}$ when the Null hypothesis represents the correlation structure correctly. And in 17 out of 36 instances in Table 9, $\hat{H}^s$ has more power when the correlation structure is misspecified than when it is correctly specified, which will likely lead the researcher to select the wrong nesting structure.

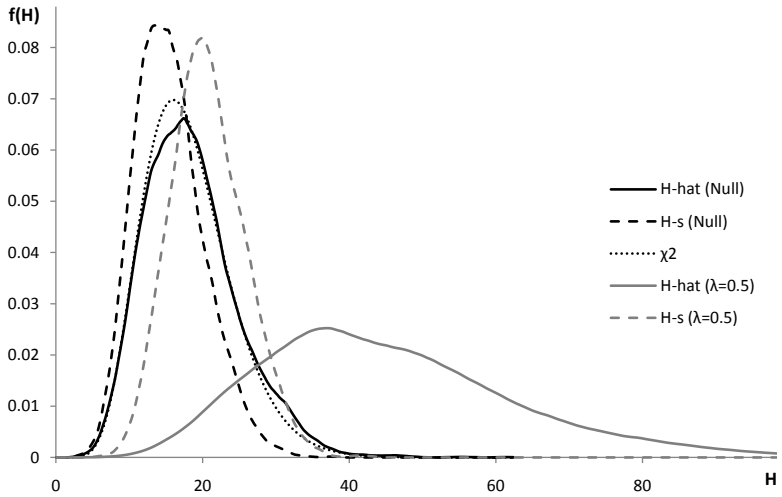


Figure 4: Comparing kernel density functions of $\hat{H}$ and $\hat{H}^s$ under Null and Alternative hypotheses (Set 3, small sample, $J = 4$, $D = \{123\}$)

The divergence between $\hat{H}$ and $\hat{H}^s$ is well demonstrated in Figure 4. Solid density curves refer to $\hat{H}$, dashed curves to $\hat{H}^s$, and the dotted curve to the χ^2 distribution. Under the Null, the gap between the density of $\hat{H}^s$ and the χ^2 curve is pronounced. When λ decreases from 1 (IIA) to 0.50, the pdf of $\hat{H}^s$ moves only gingerly to the right, whereas the pdf of $\hat{H}$ slides a large distance rightward.

Overall, therefore, the $\hat{H}^s$ test addresses the main shortcoming of the traditionally applied HM test for IIA by preventing indefinite estimated covariance matrices and negative $\tilde{H}$ outcomes, but the size distortion of $\hat{H}^s$ leads to many type-II errors in inference, unless the proper critical value is computed by simulation. Moreover, $\hat{H}^s$ is dominated by $\hat{H}$.

6 An application

The Ivorian data that are the basis for Set 3 of the simulation study offers a good contrast between the three versions of the HM test. The employment choices are: 1=farming, 2=non-farm self-employment, 3=wage employment, and 4=non-employment. Is MNL a good model to understand employment outcomes, or is the IIA assumption too restrictive?

Table 10 shows tests for IIA for every possible specification of D on the basis of $\tilde{H}$, $\hat{H}$ and $\hat{H}^s$. Five of the $\tilde{H}$ outcomes are large in the light of their χ^2 distribution; four $\tilde{H}$'s are negative; and none of the $\breve{\breve{\Omega}}_D$ matrices is positive definite. Size correction of the p-values of $\tilde{H}$ narrows the options: $D = \{234\}$ or perhaps $D = \{34\}$ looks to be a good selection, but one might wonder what the $\tilde{H}_{124}$ outcome in the far left tail signifies.

The $\hat{H}$ tests come to a different conclusion: $D = \{12\}$ draws the largest $\hat{H}$ value and the smallest p-value, rejecting IIA in favor of a nesting of outcomes 1 and 2 if the nested logit is to be adopted. A nesting of $D = \{124\}$ might also be considered.

Interestingly, the $\hat{H}^s$ tests feature two examples ($\hat{H}^s_{13}$ and $\hat{H}^s_{234}$) where the asymptotic p-value leads to non-rejection at a 5-percent significance level but the simulated p-value dictates rejection. Based on p-values, the $\hat{H}^s$ tests suggest nesting either $D = \{12\}$ or $D = \{123\}$; $\hat{H}^s_{124}$ is also statistically significant but not as large as $\hat{H}^s_{123}$.

The last column of Table 10 reports the estimates of the nesting parameter λ for each nesting structure. The nested logit model that nests outcomes 1 and 2 yields $\hat{\lambda} = 2.97$, outside the range $(0, 1)$. In fact, for nine selections of D, $\hat{\lambda}$ exceeds 1, and for the tenth ($D = \{134\}$) it equals only 0.93, too close to 1 for a meaningful nesting structure. Thus, despite the rejection of IIA by the $\hat{H}$ and $\hat{H}^s$ tests, the nested logit model may not be the right econometric model for activity choice in Cote d'Ivoire. Indeed, that the alternative hypothesis of the HM test is quite general (Train, 2009): the correct alternative model need not be a nested logit model.

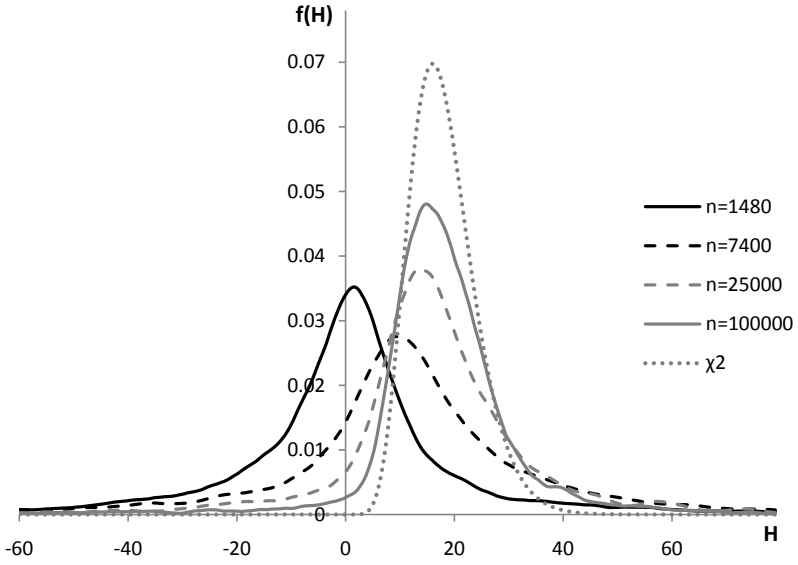


Figure 5: Kernel density functions of $\tilde{H}$ under IIA for increasing sample sizes (Set 3, $J = 4$, $D = \{123\}$)

7 Concluding remarks

The simulation and analytical results presented in this paper provide convincing evidence that the distribution of the traditional implementation of the HM test through $\tilde{H}$ deviates greatly from the asymptotic χ^2 distribution.

One might contend that the sample sizes in the Monte Carlo simulations were too small to permit substitution of asymptotic distributions. Figure 5 extends the sample size for Set 3 with $D = \{123\}$. The sampling distribution is moving towards χ^2, but for $n = 25180$ observations the goodness of fit test equals 1345 (exceeding $\chi^2_{0.99} = 36.19$), the chance of a negative $\tilde{H}$ equals 11.4 percent, and 91.3 percent of the $\breve{\breve{\Omega}}_{Dr}$ are indefinite. Even for $n = 100640$ observations, discrepancies remain large: the goodness of fit test equals 523,

the chance of a negative $\tilde{H}$ equals 5.2 percent, and 59.7 percent of the $\breve{\tilde{\Omega}}_{Dr}$ are indefinite. Meanwhile, the sampling distribution of $\hat{H}$ is indistinguishable from χ^2 for $n = 7400$ and deviates only little for $n = 1480$ (goodness of fit test equals 51).

The conclusions are clear. In its commonly applied form ($\tilde{H}$), the HM test for IIA that has been the favorite among applied researchers over the past 26 years sometimes generates negative values—or a kind warning by the software package that the covariance matrix is not positive definite. It is time to take these signals seriously. These problems arise because of an improper conceptualization of the covariance matrix and imply that, as a rule, the small-sample distribution of $\tilde{H}$ deviates dramatically from the χ^2 distribution: negative test values are common, and violations of IIA may in fact yield large *negative* test values.[20] In other words, judging the outcome of the standard HM test by the upper tail of a χ^2 distribution is likely to lead to an incorrect statistical inference. Given that there are at least 276 studies in the literature with 433 applications of the standard HM test, these findings put a significant body of empirical research at risk.

The conceptually correct implementation of the test, $\hat{H}$, is distributed approximately as χ^2 in small samples, in principle without the occurrence of negative outcomes although singularity of the covariance matrix may occasionally be an issue. $\tilde{H}$ and $\hat{H}$ are nearly uncorrelated, but other statistics that are available in a simulation study such as this strongly suggest that $\hat{H}$ is a more reliable test instrument than $\tilde{H}$. Weesie's (1999) sandwich version of the HM test, $\hat{H}^s$, offers an alternative approach to address the same shortcomings of $\tilde{H}$ but is dominated in our simulations by $\hat{H}$. Thus, if researchers want to test for IIA with the Hausman-McFadden test, they should abandon $\tilde{H}$ and use $\hat{H}$.

[20] For any given sample where it is unknown whether IIA is violated, the fact that the sampling distribution is sample-dependent implies that the density function under the null hypothesis is in fact unknown because behavior of sample members under certifiable IIA is not observed.

Appendices

A Comparing $\tilde{\Sigma}^0_{Dr}$ with Σ^0_D in large samples

Denote individuals by the double subscript it with $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Instead of letting n go to ∞, consider gathering T samples of n individuals each, one of each type i, such that $X_{it} = X_i$ for all t, with T going to ∞. This simplifying assumption is adequate for the purpose at hand. Then $T\Sigma^0_D$ equals the inverse of $-\frac{1}{T}E[h^0_D(\beta^0_D)]$. Block j, k of this expression equals

$$\frac{1}{T}E_0\left[\frac{\partial^2 \ln L_D(\beta^0_D)}{\partial\beta_j\partial\beta'_k}\right] = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} p^0_{jit}\left(I(j=k) - q^0_{kit}\right)X_{it}X'_{it} \tag{A.1}$$

which, since by assumption $X_{it} = X_i$ and thus also $p_{jit} = p_{ji}$ and $q_{kit} = q_{ki}$ for all t, is the same as equation (10).

As for $T\tilde{\Sigma}^0_{Dr}$, rewriting equation (11) in a similar way yields

$$\begin{aligned}\frac{1}{T}\frac{\partial^2 \ln \tilde{L}_D(\beta^0_D)}{\partial\beta_j\partial\beta'_k} &= -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} y_{Dit}q^0_{jit}\left(I(j=k) - q^0_{kit}\right)X_{it}X'_{it} &\text{(A.2)}\\ &= -\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T} y_{Dit}\right)q^0_{ji}\left(I(j=k) - q^0_{ki}\right)X_iX'_i. &\text{(A.3)}\end{aligned}$$

Because $\frac{1}{T}\sum_{t=1}^{T} y_{Dit}$ converges to p^0_{Di} as $T \to \infty$, and since $p^0_{Di}q^0_{ji} = p^0_{ji}$, the expression in equation (A.3) converges to that in equation (10). Thus, $T\tilde{\Sigma}^0_{Dr}$ converges to $T\Sigma^0_D$. As $\hat{\beta}$ converges to β^0 and $\breve{\beta}_D$ converges to β^0_D, $T\breve{\tilde{\Sigma}}_{Dr}$ also converges to $T\Sigma^0_D$.

B Ω_D is positive definite for any β

Without loss of generality, assume that the restricted choice set D contains the base category, and define $D^\dagger$ as the complement of D. Reorder β such that $\beta = \left(\beta'_{D^\dagger}, \beta'_D\right)'$, such that

the parameters associated with D appear at the bottom; note that $\beta_J = 0$ because of standardization. Thus, let $D^\dagger = \{1, \ldots, j_1\}$ and $D = \{j_2, \ldots, J\}$ with $j_2 = j_1 + 1$. Restate the variance of $\hat{\beta}$ as

$$\Sigma_C = \left(-E\left[\frac{\partial^2 \ln L_C}{\partial \beta \partial \beta'} \right] \right)^{-1} = \begin{pmatrix} -A_{11} & -A_{12} \\ -A_{21} & -A_{22} \end{pmatrix}^{-1} \tag{B.1}$$

which means that $\Sigma_{C,DD} = \left(-A_{22} + A_{21} A_{11}^{-1} A_{12} \right)^{-1}$. Define $\Pi_j = \text{diag}(p_{ji}^{0.5})$, $\Theta_1 = (\Pi_1 \ \ldots \ \Pi_{j_1})'$, $\Theta_2 = (\Pi_{j_2} \ \ldots \ \Pi_{J-1})'$, $\Pi_D = \text{diag}(p_{Di}^{0.5})$ and $\Pi_j^* = \Pi_j \Pi_D^{-1}$. Then, let $Z_1 = \text{diag}(\Theta_1)\,(I_{j_1} \otimes X)$ and $Z_2 = \text{diag}(\Theta_2)\,(I_{J-j_2} \otimes X)$. Given equation (4) of Section 4.1, this allows us to write $A_{11} = Z_1 \left(\Theta_1 \Theta_1' - I_{nj_1} \right) Z_1$, $A_{12} = A_{21}' = Z_1 \Theta_1 \Theta_2' Z_2$, and $A_{22} = Z_2 \left(\Theta_2 \Theta_2' - I_{n(J-j_2)} \right) Z_2$.

Furthermore, restate the variance of $\breve{\beta}_D$ as

$$\Sigma_D = \left(-E\left[\frac{\partial^2 \ln L_D}{\partial \beta_D \partial \beta_D'} \right] \right)^{-1} = -\breve{A}_{22}^{-1} \tag{B.2}$$

with block j, k of $\breve{A}_{22}$ equal to:

$$\breve{A}_{22,jk} = E\left[\frac{\partial^2 \ln L_D}{\partial \beta_j \partial \beta_k'} \right] = \sum_{i=1}^{n} \left(\frac{p_{ji} p_{ki}}{p_{Di}} - p_{ji} I(j = k) \right) X_i X_i' \tag{B.3}$$

which implies that $\breve{A}_{22} = Z_2' \left(\Theta_2 \Pi_D^{-2} \Theta_2' - I_{n(J-j_2)} \right) Z_2$.

Next, consider that $\Omega_D > 0$ if $\Sigma_D > \Sigma_{C,DD}$, i.e., if $\breve{A}_{22}^{-1} < \left(A_{22} - A_{21} A_{11}^{-1} A_{12} \right)^{-1}$, and thus if $\breve{A}_{22} - A_{22} + A_{21} A_{11}^{-1} A_{12} > 0$. From the definitions above, we have $\breve{A}_{22} - A_{22} = Z_2' \Theta_2 \left(\Pi_D^{-2} - I_n \right) \Theta_2' Z_2$. As for the other term, write $I_{nj_1} - \Theta_1 \Theta_1' = V$, such that

$$\begin{aligned} A_{21} A_{11}^{-1} A_{12} &= -Z_2' \Theta_2 \Theta_1' Z_1 \left(Z_1' V Z_1 \right)^{-1} Z_1' \Theta_1 \Theta_2' Z_2 \\ &= -Z_2' \Theta_2 \Theta_1' V^{-0.5} \left(V^{0.5} Z_1 \left(Z_1' V Z_1 \right)^{-1} Z_1' V^{0.5} \right) V^{-0.5} \Theta_1 \Theta_2' Z_2 \end{aligned} \tag{B.4}$$

As it happens, V is a symmetric $nj_1 \times nj_1$ matrix that counts $\Theta_1 (\Theta_1'\Theta_1)^{-1}$ among its eigenvectors, with eigenvalues $\Lambda_1 = (\Theta_1'\Theta_1)^{-1} - I_n$. From the definition of Θ_1, it follows that $\Theta_1'\Theta_1 = \text{diag}\left(\sum_{j=1}^{j_1} p_{ji}\right) = \text{diag}(1 - p_{Di}) = I_n - \Pi_D^2$ and therefore $\Lambda_1 = \Pi_D^2 (I_n - \Pi_D^2)^{-1}$. Moreover, $V^{-0.5}\Theta_1 = \Theta_1 (I_n - \Pi_D^2)^{0.5} \Pi_D^{-1} \equiv \Theta_1\Lambda_1^*$. Substitution of these results into equation (B.4) yields

$$\breve{A}_{22} - A_{22} + A_{21}A_{11}^{-1}A_{12} = Z_2'\Theta_2 \left(\Pi_D^{-2} - I_n\right) \Theta_2'Z_2 - Z_2'\Theta_2\Lambda_1^*\Theta_1'\Theta_1\Lambda_1^*\Theta_2'Z_2 + Z_2'\Theta_2\Lambda_1^*\Theta_1' \left(I_{nj_1} - V^{0.5}Z_1 (Z_1'VZ_1)^{-1} Z_1'V^{0.5}\right) \Theta_1\Lambda_1^*\Theta_2'Z_2 \quad \text{(B.5)}$$

The second line of equation (B.5) is a quadratic form around a idempotent matrix and is therefore positive semidefinite. With the insertion of $\Theta_1'\Theta_1 = I_n - \Pi_D^2$, the first line simplifies to $Z_2'\Theta_2 (I_n - \Pi_D^2) \Theta_2'Z_2$, which is positive definite. Thus, Ω_D is positive definite, and this is true for any β.

C $\hat{\beta}$ under conditional sampling

With conditional sampling, the properties of $\hat{\beta}$ are assessed under the assumption that the randomness of $\hat{\beta}$ arises from assigning draws of ϵ_{ji} to all members of sample S_r such that members of subsample $S_r(D)$ will always again choose an outcome from D and members of subsample $S_r(D^\dagger)$ will always again choose an outcome from $D^\dagger$. We refer to this as sampling strategy r. In other words, this strategy is dictated by the outcome of the once-drawn random sample of a given research project.

Maximizing $\ln L_C$ in equation (2) yields the first order condition

$$g_C(\hat{\beta}) = \sum_{i=1}^{n} \sum_{j \in C} (y_{ji} - \hat{p}_{ji}) X_i \quad \text{(C.1)}$$

where probability p_{ji} is evaluated at $\hat{\beta}$. Take a Taylor expansion around β_0:

$$\left(\hat{\beta} - \beta^0\right) = -h_C(\beta^0)^{-1} g_C(\beta_0) \tag{C.2}$$

where $h_C(\beta^0)$ is given in equation (4). Under conditional sampling with sampling strategy r, we have:

$$E\left[y_{ji}|r\right] = \begin{cases} \frac{p_{ji}}{1-p_{Di}} & \text{for } j \in D^\dagger \text{ and } i \in S_r(D^\dagger) \\ \frac{p_{ji}}{p_{Di}} & \text{for } j \in D \text{ and } i \in S_r(D) \\ 0 & \text{otherwise} \end{cases}$$

With this, the conditional mean of $g_C(\beta_0)$ under sampling strategy r equals

$$E\left[g_C(\beta)|r\right] = \sum_{i=1}^{n} \sum_{j \in C} a_{ji} p_{ji} X_i \tag{C.3}$$

where

$$ay_{ji} = \begin{cases} \frac{p_{ji}}{1-p_{Di}} & \text{for } j \in D^\dagger \text{ and } i \in S_r(D^\dagger) \\ \frac{p_{ji}}{p_{Di}} & \text{for } j \in D \text{ and } i \in S_r(D) \\ -1 & \text{otherwise} \end{cases}$$

Unconditionally, $E\left[g_C(\beta)\right]$ equals 0, but conditional on sampling strategy r, $E\left[g_C(\beta)|r\right]$ does not equal 0 unless by coincidence. Thus, $E\left[\hat{\beta} - \beta^0|r)\right] \neq 0$.

Next, consider the question of whether $\hat{\beta}$ is consistent. As in Appendix A, let there be n types of individuals; draw T individuals of each type to constitute a enlarged sample; and denote individuals by the double subscript it with $i = 1, \ldots, n$ and $t = 1, \ldots, T$. To examine consistency, gather T samples of n individuals each, one of each type i, such that $X_{it} = X_i$ for all t, with T going to ∞, with the condition that the proportion of the sample that chooses an outcome from set D remains the same as in the original sample S_r. In this original sample, let n_{1r} be the number of members of subsample $S_r(D)$. Let $N = nT$ be the size of the enlarged sample, and let N_1 be the number of members of the enlarged sample

that chooses $j \in D$. Then N_1/N must remain equal to n_{1r} as $T \to \infty$. Since the probability of a person of type i choosing $j \in D$ equals p^0_{Di}, the expected proportion of sample members choosing $j \in D$ equals $p^0 = \frac{1}{n}\sum_{i=1}^{n} p^0_{Di}$ where p^0_{Di} is evaluated at β^0. This is true also in the enlarged sample as $T \to \infty$. Unconditionally, N_1/N converges to p^0, but with the condition of sampling strategy r, N_1/N must equal n_{1r}. Thus, with $\hat{p}_{Di}$ evaluated at $\hat{\beta}$ as $T \to \infty$, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{Di}$ must equals n_{1r}. Unless by chance we have $n_{1r} = p^0$ in the original sample, $\hat{\beta}$ has no chance to converge to β^0 as $T \to \infty$. Thus, generally, $\hat{\beta}$ is inconsistent under sampling strategy r.

D Tests for normality

In a test of joint normality, three features of the simulated parameter estimates matter: is their mean close to the theoretical mean; is their covariance matrix close to the theoretical variance, and is the shape of the distribution close to normality? Table D1 examines the simulated parameter values for each set for $J = 3$ and $J = 4$, for the estimator of the MNL model with the overall choice set C ($\hat{\beta}$) and with one or two versions of a restricted choice set D ($\breve{\beta}_D$) and for the difference $\hat{\delta}_D$ that is the immediate focus of the Hausman-McFadden test.

Bias in the mean and variance is tested by means of a likelihood ratio test. For example, if the draws $\hat{\beta}_r$ are distributed $N(\beta^0, \Sigma^0_C)$, one may "estimate" β^0 and Σ^0_C by maximum likelihood from the simulated sample $\{\hat{\beta}_1, \ldots, \hat{\beta}_R\}$ and then observe by a likelihood ratio test whether the true β^0 and Σ^0_C fairly represent the mean and variance of $\hat{\beta}$. In Table D1, LR(mean) inserts the covariance among the simulated sample for Σ^0_C and tests whether $E[\hat{\beta}_r] = \beta^0$. LR(var) inserts the mean of the simulated sample for β^0 and tests whether the simulated covariance equals Σ^0_C. The bias in the mean is usually more pronounced than the bias in the variance, but especially for $\hat{\delta}_D$ the variance of the simulated draws deviates greatly from Ω_D.

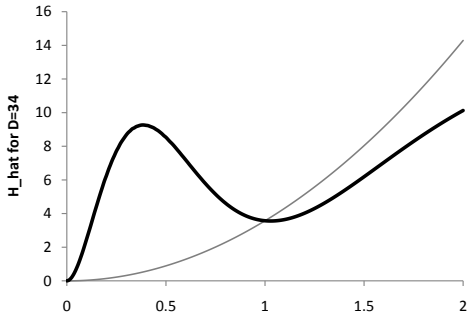
The shape of the distribution is examined with the test designed by Doornik and Hansen (2008), which considers skewness and kurtosis of each element of the parameter vector through a joint test statistic. Virtually without fail, joint normality is rejected. For elements of $\hat{\delta}_D$, while skewness is sometimes to the left and other times to the right, every element is more peaked than normality, which also means that tails are more extended.

E Singularity of $\hat{\Omega}_D$

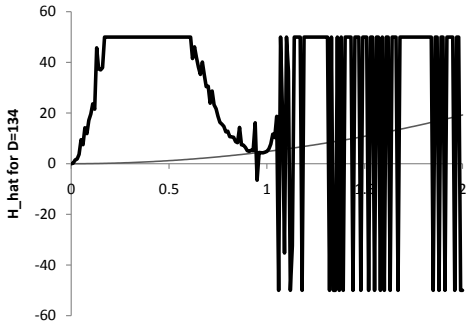
As shown in Appendix B, Ω_D is a positive definite matrix, regardless of the value of β for which Ω_D is evaluated. Nevertheless, during the simulations, a few negative values showed up for $\hat{H}$. Mathematically, this is impossible, but one would have to suspect that numerical errors are to blame for this.

Let us consider this idea. First of all, one would think that $\hat{H}$ is more or less quadratic in $\hat{\delta}_D = \breve{\beta}_D - \hat{\beta}_D$, but $\hat{\Omega}_D$ is of course a complicated function of $\hat{\beta}$. Let us therefore examine the values of $\hat{H}$ along a ray from β^0 through $\hat{\beta}$ and $\breve{\beta}_D$: see Figure E.1. Thus, we evaluate $\hat{H}$ at $\acute{\beta} = (1-\theta)\beta^0 + \theta\hat{\beta}$, where θ varies from 0 to 2: for $\theta = 1$, the Monte Carlo value obtains, and for $\theta = 0$, we have $\hat{H} = 0$. For this we select two of the 5000 random samples of the Monte Carlo analysis for Set 1, small sample, with $J = 4$. Sample r_1 yields $\hat{H}_{34} = 3.57$ and $\hat{H}_{134} = 4.81$, both of which are less than the 5-percent asymptotic critical value ($\chi^2_{0.95}(3) = 7.81$ and $\chi^2_{0.95}(6) = 12.59$. Figure E.1 shows a quadratic reference line through the Monte Carlo value at $\theta = 1$, which holds $\hat{\Omega}_D$ fixed. But clearly, $\hat{H}$ does not behave quadratically. In fact, if β were estimated closer to β^0, at around $\theta = 0.45$, $\hat{H}_{34}$ would be judged statistically significant. Moreover, the small value of $\hat{H}_{134}$ at $\theta = 1$ seems fortuitous: the behavior of $\hat{H}_{134}$ is highly erratic, to say the least.

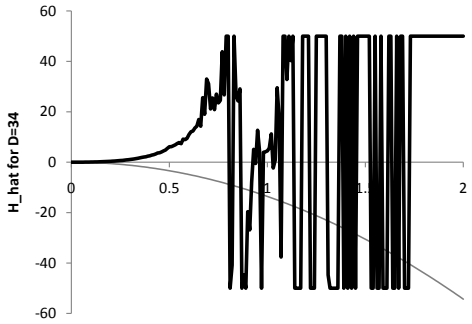
For sample r_2, the Monte Carlo values equal $\hat{H}_{34} = -13.09$ and $\hat{H}_{134} = 1.31$: along the ray, both statistics behave erratically. For $\hat{H}_{34}$, $|\hat{\Omega}_D|$ drops below 10^{-18} at $\theta = 1$, and for



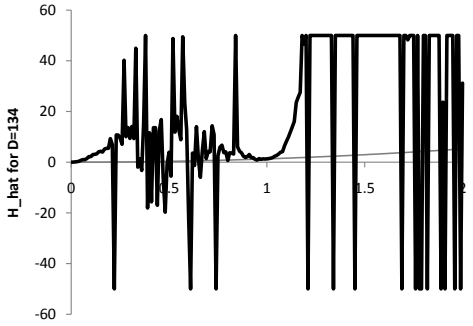
(a) Estimate r_1: $D = 34$



(b) Estimate r_1: $D = 134$



(c) Estimate r_2: $D = 34$



(d) Estimate r_2: $D = 134$

Figure E.1: $\hat{H}$ along a ray from β^0 to $2\hat{\beta}$ for two sets of estimates of $\hat{\beta}$

.

Note: The position on the ray is indicated by θ on the horizontal axis. Along the ray, $\acute{\beta} = (1-\theta)\beta^0 + \theta\hat{\beta}$. The gray curve uses $\Omega(\hat{\beta})$, whereas the black line uses $\Omega(\acute{\beta})$. Values of $\hat{H}$ are truncated at 50 and -50 to preserve the scale of the graphs.

$\hat{H}_{134}$, $|\hat{\Omega}_D|$ is always smaller than 10^{-35}.[21] Clearly, singularity of $\hat{\Omega}_D$ is playing a role.

References

Cheng, S., and J. S. Long (2007): "Testing for IIA in the Multinomial Logit Model," *Sociological Methods and Research*, 35(4), 583–600.

Doornik, J. A., and H. Hansen (2008): "An omnibus test for univariate and multivariate normality," *Oxford Bulletin of Economics and Statistics*, 70(Supp.1), 927–939.

Fry, T. R. L., and M. N. Harris (1996): "A Monte Carlo study of tests for the Independence of Irrelevant Alternatives property," *Transportation Research Part B: Methodological*, 30(1), 19–30.

——— (1998): "Testing for Independence of Irrelevant Alternatives: some empirical results," *Sociological Methods and Research*, 26(3), 401–423.

Hausman, J. (1978): "Specification Tests in econometrics," *Econometrica*, 46(6), 1251–1271.

Hausman, J., and D. McFadden (1984): "Specification Tests for the Multinomial Logit Model," *Econometrica*, 52(5), 1219–1240.

Imhof, J. P. (1961): "Computing the distribution of quadratic forms in normal variables," *Biometrika*, 48(3/4), 419–426.

McFadden, D. (1987): "Regression-based specification tests for the Multinomial Logit model," *Journal of Econometrics*, 34(1-2), 63–82.

McFadden, D., K. Train, and W. Tye (1977): "An application of diagnostic tests for the Independence of Irrelevant Alternatives property of the Multinomial Logit model," *Transportation Research Record*, 637, 39–46.

Small, K. A. (1994): "Approximate generalized extreme value models of discrete choice," *Journal of Econometrics*, 62(2), 351–382.

Small, K. A., and C. Hsiao (1985): "Multinomial Logit specification tests," *International Economic Review*, 26(3), 619–627.

Train, K. E. (2009): *Discrete Choice Methods with Simulation*. New York: Cambridge University Press, 2 edn.

[21] The careful observer may note that $\hat{H}_{34}$ along the curve does not pass through the Monte Carlo value of -13.09 but rather equals 4.47. The curve is generated with a grid search of width 0.01: the mere recalculation of $\hat{\Omega}_D$ at a value of $\hat{\beta}$ that might differ in the sixth or seventh decimal already generates a different outcome.

Vijverberg, W. P. M. (1993): "Educational investments and returns for women and men in Côte d'Ivoire," *Journal of Human Resources*, 28(4), 933–974.

Weesie, J. (1999): "Seemingly unrelated estimation and the cluster-adjusted sandwich estimator," *Stata Technical Bulletin*, 52, 34–47.

Zhang, J., and S. D. Hoffman (1993): "Discrete-Choice Logit Models: Testing the IIA property," *Sociological Methods & Research*, 22(2), 193–213.

Zheng, X. (2008): "Testing for discrete choice models," *Economics Letters*, 98(2), 176–184.

Table 1. Use of the Hausman-McFadden test in the literature, 1984-2010

	1985-2004		2005-2010		Total	
	Number	Percent	Number	Percent	Number	Percent
A: Number of tested models						
Reports all positive test values						
Fails to reject IIA	79	35.4	69	32.9	148	34.2
Rejects IIA	34	15.2	42	20.0	76	17.6
Total	113	50.7	111	52.9	224	51.7
Reports at least one negative test value						
Fails to reject IIA	23	10.3	19	9.0	42	9.7
Rejects IIA	5	2.2	2	1.0	7	1.6
Draws no conclusion	7	3.1	0	0.0	7	1.6
Use modified HM test statistic	9	4.0	5	2.4	14	3.2
Total	44	19.7	26	12.4	70	16.2
Describes test outcome verbally						
Fails to reject IIA	48	21.5	55	26.2	103	23.8
Rejects IIA	18	8.1	18	8.6	36	8.3
Total	66	29.6	73	34.8	139	32.1
Total	223	100.0	210	100.0	433	100.0
B: Number of studies that cite Hausman and McFadden (1984)						
Implementing the HM test	136		140		276	
Citing for its theoretical/ conceptual contribution	34		13		47	
Citing without apparent theoretical/conceptual contribution	45		16		61	
Incorrectly included in Web of Science:	3		1		4	
Total number of studies	218		170		388	

Source: Web of Science, accessed in March-June 2005 and February 2010, in a search for studies that cite Hausman and McFadden (1984). One study in 1992 is added that cited an earlier working paper.

Table 2. Characteristics of Monte Carlo Data Sets

	$J = 3$				$J = 4$			
	Min	Mean	Max	St.Dev.	Min	Mean	Max	St.Dev.
Set 1 ($N = 1000$)								
p_1	0.234	0.614	0.886	0.118	0.211	0.565	0.853	0.120
p_2	0.039	0.246	0.643	0.106	0.037	0.225	0.593	0.097
p_3	0.051	0.140	0.322	0.041	0.015	0.084	0.304	0.042
p_4	...	...	...	...	0.050	0.127	0.225	0.030
Set 2 ($N = 1000$)								
p_1	0.397	0.519	0.961	0.090	0.112	0.328	0.958	0.130
p_2	0.001	0.117	0.284	0.048	0.001	0.071	0.229	0.035
p_3	0.038	0.364	0.529	0.069	0.003	0.387	0.719	0.129
p_4	...	...	...	...	0.038	0.215	0.236	0.020
Set 3 ($N = 1118$ for $J = 3$, $N = 1480$ for $J = 4$)								
p_1	0.004	0.233	0.872	0.202	0.004	0.176	0.722	0.162
p_2	0.008	0.275	0.947	0.219	0.008	0.208	0.927	0.193
p_3	0.020	0.491	0.983	0.291	0.015	0.371	0.971	0.286
p_4	...	...	...	...	0.004	0.245	0.831	0.265

Table 3: Behavior of H under the Null hypothesis

		Small sample[a]				Large sample[b]			
		Empirical size at a nominal size of			Goodness of fit	Empirical size at a nominal size of			Goodness of fit
	D	0.10	0.05	0.01	$\chi^2(19)$[c]	0.10	0.05	0.01	$\chi^2(19)$[c]
A: $J = 3$									
Set 1	12	0.153	0.103	0.049	343	0.106	0.053	0.010	18
	13	0.368	0.309	0.221	7168	0.180	0.118	0.051	536
	23	0.323	0.268	0.178	5130	0.154	0.100	0.037	301
Set 2	12	0.263	0.202	0.126	2495	0.136	0.083	0.028	129
	13	0.277	0.221	0.147	3241	0.137	0.087	0.035	184
	23	0.406	0.346	0.250	9395	0.179	0.123	0.054	607
Set 3	12	0.224	0.152	0.068	1216	0.113	0.063	0.015	35
	13	0.203	0.138	0.062	902	0.118	0.060	0.015	36
	23	0.193	0.129	0.051	746	0.113	0.062	0.012	31
B: $J = 4$									
Set 1	12	0.166	0.112	0.053	432	0.105	0.056	0.014	25
	123	0.366	0.298	0.200	6672	0.149	0.091	0.031	230
	234	0.872	0.852	0.814	67808	0.586	0.533	0.432	24659
Set 2	12	0.326	0.269	0.188	5164	0.151	0.094	0.037	231
	123	0.391	0.325	0.232	8143	0.165	0.099	0.036	329
	234	0.596	0.539	0.426	25373	0.241	0.172	0.089	1677
Set 3	12	0.232	0.162	0.080	1474	0.126	0.066	0.017	68
	123	0.264	0.190	0.089	2271	0.136	0.077	0.024	143
	234	0.267	0.186	0.090	2236	0.138	0.082	0.023	145

[a] $N = 1000$ for Sets 1 and 2; for Set 3, $N = 1118$ for $J = 3$ and $N = 1480$ for $J = 4$,
[b] $N = 7500$ for Sets 1 and 2; for Set 3, $N = 7826$ for $J = 3$ and $N = 7400$ for $J = 4$.
[c] Critical values are 27.20 (10 percent), 30.14 (5 percent) and 36.19 (1 percent).

Table 4: Behavior of $\tilde{H}$ under the Null hypothesis

		Empirical size at a nominal size of			Indefinite $\breve{\tilde{\Omega}}_{Dr}$			Goodness of fit
	D	0.10	0.05	0.01	$\tilde{H} \leq 0$	$\tilde{H} > 0$	Total	$\chi^2(19)$[a]
A1: $J = 3$, *Small sample*[b]								
Set 1	12	0.045	0.039	0.027	0.408	0.352	0.760	3918
	13	0.023	0.020	0.012	0.407	0.400	0.807	8357
	23	0.044	0.035	0.024	0.380	0.323	0.703	4082
Set 2	12	0.041	0.034	0.027	0.344	0.572	0.916	9766
	13	0.001	0.001	0.001	0.589	0.382	0.971	16333
	23	0.033	0.026	0.020	0.395	0.513	0.908	10661
Set 3	12	0.108	0.097	0.077	0.502	0.485	0.987	4655
	13	0.168	0.147	0.114	0.401	0.530	0.930	4045
	23	0.116	0.101	0.080	0.478	0.504	0.982	4341
A2: $J = 3$, *Large sample*[c]								
Set 1	12	0.105	0.083	0.056	0.155	0.406	0.561	1304
	13	0.083	0.068	0.048	0.249	0.459	0.708	3246
	23	0.112	0.090	0.059	0.149	0.409	0.558	1375
Set 2	12	0.089	0.073	0.051	0.281	0.502	0.782	3101
	13	0.012	0.011	0.008	0.561	0.381	0.942	10208
	23	0.105	0.087	0.064	0.288	0.505	0.793	3204
Set 3	12	0.292	0.240	0.171	0.091	0.414	0.504	4227
	13	0.213	0.166	0.099	0.076	0.328	0.405	1753
	23	0.277	0.231	0.162	0.082	0.381	0.464	3863
B1: $J = 4$, *Small sample*[d]								
Set 1	12	0.066	0.054	0.038	0.384	0.362	0.746	3019
	123	0.045	0.038	0.027	0.422	0.507	0.930	11966
	234	0.037	0.029	0.022	0.284	0.593	0.877	18242
Set 2	12	0.056	0.048	0.034	0.337	0.574	0.911	7945
	123	0.059	0.051	0.040	0.473	0.506	0.979	9743
	234	0.026	0.022	0.018	0.426	0.550	0.976	18053
Set 3	12	0.153	0.135	0.109	0.428	0.533	0.961	4199
	123	0.106	0.097	0.084	0.471	0.529	1.000	9185
	234	0.085	0.077	0.062	0.563	0.437	1.000	7325
B2: $J = 4$, *Large sample*[e]								
Set 1	12	0.111	0.084	0.051	0.138	0.403	0.541	1086
	123	0.099	0.085	0.061	0.127	0.756	0.883	8995
	234	0.044	0.034	0.025	0.173	0.696	0.869	15706
Set 2	12	0.101	0.081	0.056	0.256	0.519	0.775	2652
	123	0.122	0.098	0.072	0.145	0.806	0.950	6532
	234	0.102	0.087	0.066	0.247	0.710	0.957	9372
Set 3	12	0.279	0.231	0.162	0.090	0.380	0.470	3876
	123	0.222	0.199	0.156	0.239	0.758	0.997	6635
	234	0.238	0.214	0.175	0.272	0.726	0.998	6824

[a] Critical values are 27.20 (10 percent), 30.14 (5 percent) and 36.19 (1 percent).
[b] $N = 1000$ for Sets 1 and 2; $N = 1118$ for Set 3.
[c] $N = 7500$ for Sets 1 and 2; $N = 7826$ for Set 3.
[d] $N = 1000$ for Sets 1 and 2; $N = 1480$ for Set 3.
[e] $N = 7500$ for Sets 1 and 2; $N = 7400$ for Set 3.

Table 5: Behavior of $\hat{H}$ under the Null hypothesis

		Small sample[a]				Large sample[b]			
		Empirical size at a nominal size of			Goodness of fit	Empirical size at a nominal size of			Goodness of fit
	D	0.10	0.05	0.01	$\chi^2(19)$[c]	0.10	0.05	0.01	$\chi^2(19)$[c]
A: $J = 3$									
Set 1	12	0.094	0.049	0.015	22	0.096	0.045	0.007	23
	13	0.105	0.053	0.015	15	0.101	0.053	0.011	25
	23	0.107	0.060	0.017	25	0.100	0.049	0.014	12
Set 2	12	0.099	0.053	0.016	22	0.104	0.050	0.012	19
	13	0.083	0.043	0.016	139	0.097	0.050	0.017	43
	23	0.107	0.057	0.015	28	0.102	0.049	0.011	31
Set 3	12	0.115	0.066	0.017	50	0.100	0.051	0.010	15
	13	0.122	0.066	0.017	49	0.112	0.054	0.008	25
	23	0.114	0.062	0.016	21	0.100	0.050	0.010	10
B: $J = 4$									
Set 1	12	0.099	0.053	0.013	20	0.093	0.046	0.010	18
	123	0.104	0.061	0.019	56	0.097	0.049	0.011	15
	234	0.134	0.081	0.031	128	0.098	0.054	0.011	35
Set 2	12	0.116	0.070	0.024	64	0.101	0.052	0.010	21
	123	0.115	0.067	0.020	59	0.100	0.048	0.009	14
	234	0.107	0.058	0.013	33	0.105	0.053	0.012	18
Set 3	12	0.138	0.077	0.023	104	0.106	0.052	0.012	26
	123	0.122	0.068	0.014	51	0.106	0.052	0.014	26
	234	0.115	0.062	0.016	36	0.109	0.055	0.012	17

[a] $N = 1000$ for Sets 1 and 2; for Set 3, $N = 1118$ for $J = 3$ and $N = 1480$ for $J = 4$,
[b] $N = 7500$ for Sets 1 and 2; for Set 3, $N = 7826$ for $J = 3$ and $N = 7400$ for $J = 4$.
[c] Critical values are 27.20 (10 percent), 30.14 (5 percent) and 36.19 (1 percent).

Table 6: Size-adjusted power of $\tilde{H}$ for three nesting structures

	D	Power			$\tilde{H} \leq 0$			$\breve{\breve{\Omega}}_{Dr}$ is indefinite		
		(12)3	2(13)	1(23)	(12)3	2(13)	1(23)	(12)3	2(13)	1(23)
A1: $J = 3$, *Small sample*										
Set 1	12	0.074	0.023	0.021	0.195	0.527	0.500	0.369	0.866	0.848
	13	0.074	0.170	0.190	0.441	0.210	0.198	0.867	0.764	0.764
	23	0.177	0.072	0.083	0.292	0.276	0.279	0.723	0.764	0.764
Set 2	12	0.051	0.043	0.095	0.197	0.407	0.238	0.905	0.932	0.934
	13	0.006	0.284	0.011	0.632	0.482	0.621	0.990	0.914	0.981
	23	0.096	0.043	0.070	0.238	0.468	0.283	0.912	0.910	0.861
Set 3	12	0.026	0.085	0.065	0.050	0.668	0.621	0.113	0.995	0.989
	13	0.091	0.007	0.050	0.525	0.012	0.566	0.983	0.018	0.995
	23	0.026	0.048	0.036	0.617	0.644	0.017	0.999	1.000	0.026
A2: $J = 3$, *Large sample*										
Set 1	12	0.053	0.072	0.057	0.008	0.273	0.291	0.010	0.718	0.703
	13	0.099	0.025	0.027	0.276	0.224	0.222	0.894	0.507	0.509
	23	0.140	0.039	0.045	0.116	0.223	0.216	0.828	0.713	0.704
Set 2	12	0.022	0.055	0.063	0.261	0.294	0.214	0.746	0.836	0.908
	13	0.001	0.397	0.012	0.578	0.300	0.610	0.973	0.646	0.973
	23	0.029	0.056	0.028	0.209	0.325	0.250	0.862	0.792	0.500
Set 3	12	0.000	0.456	0.353	0.000	0.428	0.355	0.000	0.999	0.966
	13	0.737	0.875	0.478	0.068	0.000	0.068	0.975	0.000	1.000
	23	0.093	0.276	0.038	0.602	0.420	0.000	0.897	1.000	0.000
		(12)34	(123)4	1(234)	(12)34	(123)4	1(234)	(12)34	(123)4	1(234)
B1: $J = 4$, *Small sample*										
Set 1	12	0.048	0.047	0.053	0.164	0.193	0.547	0.324	0.396	0.887
	123	0.065	0.070	0.098	0.226	0.155	0.396	0.840	0.845	0.990
	234	0.104	0.135	0.070	0.277	0.310	0.102	0.940	0.942	0.686
Set 2	12	0.027	0.037	0.066	0.220	0.219	0.213	0.892	0.855	0.957
	123	0.058	0.053	0.089	0.397	0.093	0.430	0.980	0.912	0.996
	234	0.067	0.176	0.120	0.353	0.428	0.206	0.990	0.996	0.943
Set 3	12	0.012	0.046	0.078	0.030	0.113	0.450	0.058	0.236	0.932
	123	0.073	0.045	0.060	0.402	0.127	0.574	1.000	0.770	1.000
	234	0.048	0.049	0.101	0.530	0.687	0.129	1.000	1.000	0.582
B2: $J = 4$, *Large sample*										
Set 1	12	0.124	0.045	0.256	0.002	0.011	0.288	0.003	0.015	0.881
	123	0.025	0.011	0.093	0.048	0.023	0.202	0.583	0.465	0.991
	234	0.284	0.294	0.412	0.155	0.136	0.042	0.993	0.981	0.416
Set 2	12	0.024	0.043	0.029	0.283	0.279	0.216	0.747	0.646	0.973
	123	0.030	0.098	0.126	0.102	0.007	0.119	0.942	0.755	0.997
	234	0.042	0.466	0.023	0.252	0.060	0.053	0.984	1.000	0.732
Set 3	12	0.001	0.001	0.159	0.000	0.000	0.081	0.000	0.001	0.851
	123	0.045	0.031	0.132	0.130	0.016	0.380	0.780	0.041	1.000
	234	0.050	0.084	0.001	0.417	0.574	0.002	1.000	1.000	0.004

Table 7: Size-adjusted power of $\hat{H}$ for various nesting structures

A: $J = 3$

		Small sample			Large sample		
	D	(12)3	2(13)	1(23)	(12)3	2(13)	1(23)
Set 1	12	0.157	0.056	0.056	0.531	0.050	0.052
	13	0.044	0.072	0.069	0.085	0.085	0.086
	23	0.036	0.059	0.055	0.185	0.054	0.055
Set 2	12	0.065	0.051	0.051	0.071	0.044	0.069
	13	0.043	0.066	0.045	0.048	0.082	0.045
	23	0.053	0.065	0.081	0.060	0.047	0.120
Set 3	12	0.510	0.067	0.068	0.998	0.959	0.678
	13	0.057	0.709	0.089	0.848	1.000	0.855
	23	0.053	0.059	0.592	0.183	0.860	1.000

B: $J = 4$

		Small sample			Large sample		
	D	(12)34	(123)4	1(234)	(12)34	(123)4	1(234)
Set 1	12	0.163	0.118	0.048	0.679	0.463	0.228
	123	0.096	0.142	0.031	0.304	0.465	0.103
	234	0.033	0.042	0.234	0.165	0.097	0.696
Set 2	12	0.064	0.091	0.040	0.081	0.342	0.049
	123	0.057	0.257	0.033	0.058	0.960	0.119
	234	0.052	0.098	0.139	0.049	0.681	0.357
Set 3	12	0.595	0.210	0.034	0.999	0.576	0.151
	123	0.118	0.816	0.045	0.379	1.000	0.381
	234	0.050	0.053	0.763	0.190	0.488	1.000

Note: For the definition of sample size, see Table 3.

Table 8: Behavior of $\hat{H}^s$ under the Null hypothesis

		Small sample[a]				Large sample[b]			
		Empirical size at a nominal size of			Goodness of fit	Empirical size at a nominal size of			Goodness of fit
	D	0.10	0.05	0.01	$\chi^2(19)$[c]	0.10	0.05	0.01	$\chi^2(19)$[c]
A: $J = 3$									
Set 1	12	0.044	0.019	0.004	380	0.077	0.036	0.007	43
	13	0.018	0.004	0.000	1267	0.053	0.022	0.003	261
	23	0.014	0.004	0.000	1500	0.049	0.023	0.003	219
Set 2	12	0.027	0.012	0.001	920	0.066	0.029	0.007	97
	13	0.006	0.002	0.000	1732	0.064	0.031	0.006	188
	23	0.006	0.000	0.000	1953	0.045	0.018	0.001	338
Set 3	12	0.025	0.007	0.000	776	0.075	0.036	0.005	57
	13	0.037	0.015	0.001	398	0.093	0.042	0.005	22
	23	0.035	0.010	0.001	468	0.080	0.038	0.007	48
B: $J = 4$									
Set 1	12	0.048	0.021	0.004	371	0.077	0.032	0.007	49
	123	0.032	0.015	0.003	1424	0.057	0.024	0.004	165
	234	0.007	0.003	0.000	7427	0.017	0.006	0.001	1972
Set 2	12	0.022	0.008	0.000	1089	0.063	0.028	0.004	158
	123	0.020	0.008	0.001	1666	0.069	0.031	0.005	136
	234	0.006	0.002	0.000	5338	0.028	0.010	0.002	727
Set 3	12	0.034	0.014	0.001	545	0.074	0.032	0.007	88
	123	0.024	0.008	0.000	918	0.081	0.036	0.006	46
	234	0.019	0.006	0.001	1207	0.075	0.034	0.006	93

[a] $N = 1000$ for Sets 1 and 2; for Set 3, $N = 1118$ for $J = 3$ and $N = 1480$ for $J = 4$,
[b] $N = 7500$ for Sets 1 and 2; for Set 3, $N = 7826$ for $J = 3$ and $N = 7400$ for $J = 4$.
[c] Critical values are 27.20 (10 percent), 30.14 (5 percent) and 36.19 (1 percent).

Table 9: Size-adjusted power of $\hat{H}^s$ for various nesting structures

A: $J = 3$

		Small sample			Large sample		
	D	(12)3	2(13)	1(23)	(12)3	2(13)	1(23)
Set 1	12	0.069	0.078	0.156	0.413	0.078	0.235
	13	0.075	0.064	0.110	0.100	0.073	0.248
	23	0.160	0.067	0.164	0.326	0.063	0.521
Set 2	12	0.045	0.056	0.064	0.055	0.058	0.116
	13	0.026	0.100	0.045	0.045	0.059	0.043
	23	0.088	0.073	0.086	0.101	0.054	0.111
Set 3	12	0.148	0.402	0.270	0.992	0.991	0.849
	13	0.282	0.210	0.215	0.925	1.000	0.909
	23	0.081	0.238	0.177	0.283	0.938	0.998

B: $J = 4$

		Small sample			Large sample		
	D	(12)34	(123)4	1(234)	(12)34	(123)4	1(234)
Set 1	12	0.072	0.061	0.161	0.562	0.334	0.361
	123	0.042	0.060	0.194	0.180	0.306	0.308
	234	0.128	0.167	0.144	0.372	0.342	0.702
Set 2	12	0.050	0.048	0.072	0.045	0.247	0.082
	123	0.050	0.093	0.134	0.036	0.884	0.270
	234	0.074	0.194	0.135	0.080	0.747	0.414
Set 3	12	0.176	0.072	0.132	0.984	0.334	0.318
	123	0.050	0.247	0.191	0.174	1.000	0.678
	234	0.084	0.181	0.240	0.374	0.724	0.999

Note: For the definition of sample size, see Table 3.

Table 10: $\tilde{H}$, $\hat{H}$, and $\hat{H}^s$ tests for IIA: Activity choice in Côte d'Ivoire

	$\tilde{H}$			$\hat{H}$			$\hat{H}^s$			Nested logit
		p-value			p-value			p-value		
D	Value	Asymp	Sim	Value	Asymp	Sim	Value	Asymp	Sim	$\hat{\lambda}$
12	-10.60	...	0.800	73.62	0.000	0.000	27.20	0.001	0.000	2.97*
13	26.33	0.002	0.135	12.49	0.187	0.216	14.94	0.092	0.050	1.60
14	-49.04	...	0.950	16.17	0.063	0.094	19.88	0.019	0.004	2.52*
23	30.11	0.000	0.095	26.66	0.002	0.003	19.14	0.024	0.006	1.71
24	39.72	0.000	0.086	13.25	0.152	0.186	13.24	0.152	0.084	1.43
34	75.70	0.000	0.041	25.37	0.003	0.009	21.63	0.010	0.003	1.03
123	-15.54	...	0.821	40.96	0.002	0.004	35.74	0.008	0.000	1.59
124	-171.44	...	0.982	55.48	0.000	0.000	32.91	0.017	0.001	2.42*
134	17.74	0.473	0.136	22.49	0.211	0.253	16.68	0.545	0.317	0.93
234	87.07	0.000	0.028	31.14	0.028	0.038	27.12	0.077	0.012	1.12

Notes: p-values evaluated at a 5% significance level

* The t-ratio of $\hat{\lambda} - 1$ exceeds 2

The matrix $\breve{\hat{\Omega}}_D$ is indefinite for every choice of D.

Table D1: P-values of tests for normality of $\hat{\beta}$, $\breve{\beta}_D$ and $\hat{\delta}_D$

	Choice set	Parameter	$J=3$ LR(mean)	$J=3$ LR(var)	$J=3$ Doornik-Hansen	$J=4$ LR(mean)	$J=4$ LR(var)	$J=4$ Doornik-Hansen
A: Small Sample								
Set 1	C	$\hat{\beta}$	$< 10^{-9}$	0.4467	$< 10^{-4}$	$< 10^{-20}$	0.0002	$< 10^{-6}$
	$D=12$	$\breve{\beta}_D$	0.0010	0.0357	$< 10^{-5}$	0.0044	0.4609	0.0071
		$\hat{\delta}_D$	0.0062	$< 10^{-100}$	$< 10^{-100}$	0.0036	$< 10^{-100}$	$< 10^{-100}$
	$D=123$	$\breve{\beta}_D$				$< 10^{-14}$	0.6836	0.0146
		$\hat{\delta}_D$				0.0022	$< 10^{-100}$	$< 10^{-100}$
Set 2	C	$\hat{\beta}$	$< 10^{-31}$	0.0026	$< 10^{-81}$	$< 10^{-73}$	$< 10^{-5}$	$< 10^{-100}$
	$D=12$	$\breve{\beta}_D$	$< 10^{-32}$	$< 10^{-5}$	$< 10^{-78}$	$< 10^{-56}$	$< 10^{-10}$	$< 10^{-100}$
		$\hat{\delta}_D$	0.0061	$< 10^{-100}$	$< 10^{-100}$	$< 10^{-6}$	$< 10^{-100}$	$< 10^{-100}$
	$D=123$	$\breve{\beta}_D$				$< 10^{-68}$	$< 10^{-7}$	$< 10^{-100}$
		$\hat{\delta}_D$				0.1352	$< 10^{-100}$	$< 10^{-100}$
Set 3	C	$\hat{\beta}$	$< 10^{-71}$	0.0045	$< 10^{-12}$	$< 10^{-100}$	0.0001	$< 10^{-26}$
	$D=12$	$\breve{\beta}_D$	$< 10^{-44}$	$< 10^{-45}$	$< 10^{-12}$	$< 10^{-50}$	$< 10^{-4}$	$< 10^{-19}$
		$\hat{\delta}_D$	$< 10^{-22}$	$< 10^{-100}$	$< 10^{-27}$	$< 10^{-23}$	$< 10^{-100}$	$< 10^{-55}$
	$D=123$	$\breve{\beta}_D$				$< 10^{-67}$	0.0009	$< 10^{-31}$
		$\hat{\delta}_D$				$< 10^{-13}$	$< 10^{-100}$	$< 10^{-38}$
B: Large Sample								
Set 1	C	$\hat{\beta}$	0.0297	0.7034	0.1836	0.0009	0.0002	0.0821
	$D=12$	$\breve{\beta}_D$	0.6309	0.2942	0.3944	0.0521	0.2169	0.0479
		$\hat{\delta}_D$	0.1008	0.0007	$< 10^{-16}$	0.8187	0.0309	$< 10^{-5}$
	$D=123$	$\breve{\beta}_D$				0.0119	0.6481	0.4612
		$\hat{\delta}_D$				0.8904	$< 10^{-97}$	$< 10^{-28}$
Set 2	C	$\hat{\beta}$	$< 10^{-5}$	0.4214	$< 10^{-8}$	$< 10^{-8}$	0.0239	$< 10^{-11}$
	$D=12$	$\breve{\beta}_D$	$< 10^{-6}$	0.0197	$< 10^{-7}$	$< 10^{-8}$	0.5356	$< 10^{-15}$
		$\hat{\delta}_D$	0.0079	$< 10^{-48}$	$< 10^{-17}$	0.3324	$< 10^{-100}$	$< 10^{-32}$
	$D=123$	$\breve{\beta}_D$				$< 10^{-}8$	0.0430	$< 10^{-}11$
		$\hat{\delta}_D$				0.5370	$< 10^{-100}$	$< 10^{-34}$
Set 3	C	$\hat{\beta}$	$< 10^{-9}$	0.5072	0.0075	$< 10^{-18}$	0.7662	$< 10^{-8}$
	$D=12$	$\breve{\beta}_D$	$< 10^{-5}$	$< 10^{-20}$	0.0339	$< 10^{-11}$	0.7286	0.0035
		$\hat{\delta}_D$	0.1475	$< 10^{-6}$	0.1174	$< 10^{-10}$	$< 10^{-8}$	0.3267
	$D=123$	$\breve{\beta}_D$				$< 10^{-14}$	0.9585	$< 10^{-7}$
		$\hat{\delta}_D$				0.0020	$< 10^{-20}$	0.0010

Notes: LR(mean) refers to a joint likelihood ratio test on the means of the parameter estimator, given their covariance and assuming normality.
LR(variance) refers to a joint likelihood ratio test on the covariance matrix of the parameter estimator, given their mean and assuming normality.
The Doornik-Hansen test examines skewness and kurtosis of all parameter estimators jointly.