Transportation Research Part B 114 (2018) 241–253

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Overidentification tests for the exogeneity of instruments in discrete choice models

C. Angelo Guevara

Departamento de Ingeniería Civil, Universidad de Chile, Facultad de Ingeniería y Ciencias Aplicadas, Blanco Encalada 2002, Santiago, Chile

ARTICLE INFO

Article history:
Received 17 March 2018
Revised 27 April 2018
Accepted 30 May 2018
Available online 19 June 2018

Keywords:
Endogeneity
Overidentification
Control-function
Logit
Probit
Instrumental variables

ABSTRACT

Endogeneity is often present in discrete choice models, precluding the consistent estimation of the model parameters. To correct for this problem, the researcher needs to gather exogenous instrumental variables, which should be independent of the error term of the model. This critical assumption can be tested using overidentification tests that rely on having more instruments than endogenous variables. For discrete choice models, instruments' exogeneity can be assessed using the Amemiya-Lee-Newey test, which relies in the estimation of an auxiliary GMM model build from reduced-form estimates. This paper proposes two alternative tests that are constructed as adaptations of the Refutability test and the Hausman test, into the discrete choice framework. The Refutability test consists in including instruments as additional variables in an auxiliary model that was corrected for endogeneity using all instruments available. The Hausman test is built from the comparison of the estimates attained using different subsets of instruments. Using a binary choice Monte Carlo experiment, the three tests are assessed in terms of size, power, and robustness to De Blander's condition for which all overidentification tests of this kind are blind. Results show that what is termed the modified Refutability test, which includes all instruments simultaneously, has smaller size distortion, larger power, and more robustness, compared to the state of the art Amemiya-Lee-Newey test and to the proposed Hausman test. Besides, the Amemiya-Lee-Newey and both versions of the Refutability test allow being agnostic about which instrument might be exogenous. The paper finishes highlighting the methodological and practical implications and limitations of these findings and suggesting future lines of research in this area.

1. Introduction

When the error term of a discrete choice model is not independent of the observed variables, conventional estimators of the model parameters are inconsistent, making the model misleading for behavioral assessment and policy design (see e.g. Guevara and Thomas, 2007). This problem is known as endogeneity and may be caused by three main reasons: errors in variables, simultaneous determination and the omission of attributes of alternatives (Guevara, 2015). Endogeneity is common to several types of discrete choice models used in transportation analysis, including, but not limited to the following: airline itinerary choice (Lurkin et al., 2017), mode choice (Fernandez-Antolin et al, 2016), passenger booking timing (Wen and Chen, 2017), learning models of route choice (Guevara et al., 2017), mobility data collection (Zegras et al., 2018),

E-mail address: crguevar@ing.uchile.cl

https://doi.org/10.1016/j.trb.2018.05.020
0191-2615/© 2018 Elsevier Ltd. All rights reserved.

demand for electric vehicles (Helveston, 2016), vehicles' purchases (Petrin and Train, 2010), valuation of public transport attributes (Guevara et al., 2018) and residential choice (Guevara 2005, 2010; Guevara and Ben-Akiva 2006, 2012; Ferreira, 2010; Guevara and Polanco, 2016).

The canonical methods to correct for endogeneity in discrete choice models rely on the availability of suitable instrumental variables that must comply with two seemingly opposing properties: relevance and exogeneity. On the one hand instruments must be correlated with the endogenous variable (relevant) and on the other they must be independent of the error term of the model (exogenous). Therefore, finding suitable instruments in practice can be challenging and even polemic (see e.g. Hausman, 1997; Bresnahan, 1997).

Mumbower et al. (2014) reviewed different types of possible sources for instrumental variables that have been proposed in the literature, classifying them into four categories. The first category (cost-shifting) corresponds to instruments that are based on variables that share marginal costs with the endogenous variable, but that differ in demand shocks that explain the error term. The second category (Stern-type) uses instruments that are built from measurable differences in market power positions among markets. The third category (Hausman-type) takes advantage of differences in geographical contexts to achieve the cost-shifting goal. The final category (BLP-type) corresponds to instruments that are based on measures of non-price characteristics of other products supplied by the same firm.

For example, Guevara and Ben-Akiva (2006, 2012) use what can be classified as Hausman type instruments to address a price endogeneity problem in a residential location choice model. The source of endogeneity in this case was that dwelling price was likely correlated with omitted attributes. The problem was addressed by the authors using the prices of other dwellings located not too far but neither too close to the incumbent dwelling. By this, relevance would be achieved because close dwellings may share marginal cost, and exogeneity would hold because dwellings are located beyond some limit that permits assuming that they differ in the demand shocks that may partially explain the error term of the model.

In practice, instrumental variables must be gathered based on the researcher's judgment on the validity of the proposed source for cost-shifting, Stern, Hausman or BLP instruments. However, without formal tests, the suitability of the proposed instruments is always arguable. Testing if the instruments are relevant could be achieved by analyzing the correlation between the instrument and the endogenous variable, which are both observable. The challenge in such a case is to determine a statistic that can guarantee, with certain confidence, that the instruments are sufficiently correlated with the endogenous variable to attain certain goal in bias or empirical coverage (Guevara and Navarro, 2015). Testing instruments exogeneity is somehow more difficult because the error term of the model is latent (not observed), hindering the construction of a statistic to judge their independence from the instrumental variables. This article will tackle the latter challenge in the context of discrete choice models.

The challenge of judging the independence between the instrumental variable and the latent error for assessing exogeneity, is achieved by relying on overidentification. To correct for endogeneity, only one instrument per endogenous variable is needed. If more than the minimum needed instruments are used, the corrected model is said to be overidentified. If all instruments are exogenous, estimators will be consistent, and then, e.g., any difference attained with alternative sets of instruments used could only be attributed to sampling error, which allows testing instruments' validity. For linear models, Sargan (1958) noted that when the problem is overidentified the residuals of the instrumental-variables regression can be used to test for the exogeneity of the instruments. For nonlinear models, including discrete choice models such as the logit or the probit, Lee (1992) noted that an estimator developed by Amemiya (1978), and studied by Newey (1987), can play the role of the Sargan test. This test is usually termed as the Amemiya-Lee-Newey (ALN) test, it relies on the estimation of an auxiliary GMM model build from reduced-form estimates, and it is the state of the art in the subject.

Overidentification tests for the exogeneity of the instruments have a key limitation. Newey (1985) showed that these tests are inconsistent, which means that they are blind to certain alternative hypothesis even if the sample size goes to infinity. One way to recover consistency is to consider that the overidentification tests work under the assumption that a subset of the instruments, for which the model becomes just identified, is exogenous (Stock, 2001). This additional assumption cannot be proven, what seems to discourage the use of methods to correct for endogeneity that are based on instrumental variables.

However, De Blander (2008) proposes an alternative way to attain consistency of overidentification tests, which seems to increase their practical appealing. He notes that the alternative hypothesis for which overidentification tests are blind is very peculiar, so he recommends instead to assume that the conditions that produce this alternative hypothesis do not hold. This change puts "the burden of proof … on the critic, who has to make the case why the instruments" would fulfill this rare condition. De Blander (2008) shows that consistency would fail if the way in which the instruments appear in the structural equation and the reduced form equation, are linearly dependent. Pleus (2015) provides a more general expression for this result, building on Newey (1985), and presents a graphical representation to illustrate the nature of the problem. Parente and Silva (2012) identify one plausible case in which this peculiar condition may occur in practice, which is when both instruments are of the same nature, if they come from the same source. The problem is that, in such a case, the correlation of the instruments with both the endogenous variable and the error term will likely be very similar. An analogous warning was previously recommended by Nichols (2007), although justified on a different ground. Beyond Parente and Silva's (2012) warning, which is fully addressable in practice, it seems then easier to defend than to attack the plausibility of the consistency of the overidentification tests for the validity of instruments.

This article focuses on the development and the assessment of tests for the exogeneity of instruments into the discrete choice modeling framework. First, the state of the art Amemiya-Lee-Newey test is reviewed and compared with two novel tests that are constructed as adaptations of the Refutability and the Hausman test into the discrete choice framework. In the Refutability test, instruments are included as additional variables in an auxiliary model that was corrected for endogeneity. An early version of this test, in which not all instruments are included in the auxiliary model, was suggested by Guevara (2010), who termed it the Direct test. Here is presented also a modified version of the Refutability test that allows evaluating all instruments simultaneously. The Hausman test is built from the comparison of the estimates attained using all and a subset of instruments in the correction of endogeneity, taking advantage of the differences on efficiency that would be attained for each case.

The three types of tests are assessed, using a binary choice Monte Carlo experiment, in terms of size, power and their robustness to De Blander's condition to which all overidentification tests of this kind are blind. Results suggests that the modified version of the Refutability test has larger power, smaller size distortion, and is more robust, compared to the state of the art Amemiya-Lee-Newey test. Besides, the Amemiya-Lee-Newey and both versions of the Refutability test allow being agnostic about which instrument might be exogenous.

The paper is structured as follows. Section 2 deploys a reference choice model that suffers from endogeneity. This reference model is used in the next sections as the testbed on which the proposed tests and the Monte Carlo experiments are developed. Sections 3, 4 and 5 convey the formulation of the three overidentification tests for the exogeneity of instruments in discrete choice models, beginning with the state of the art Amemiya-Lee-Newey, followed by the proposed Refutability and then the Hausman tests. Section 6 reports the Monte Carlo experiment. The paper finishes highlighting the practical implications and limitations of these findings, and suggesting future lines of research in this area.

2. Reference discrete choice model

For illustrative purposes, consider the reference model depicted in Eqs. (1)–(3) where endogeneity is present and can be addressed using instrumental variables. The model is simple enough for clarity of exposition, and is motivated as a discrete choice model. However, it can be easily extended or re-interpreted to represent a vast variety of problems that involve limited dependent variables (Wooldridge, 2010) that are relevant for transportation science.

$$U^*_{in} = \beta_i + \beta_p p_{in} + \beta_x x_{in} + \varepsilon^*_{in} \tag{1}$$

$$p_{in} = \alpha_0 + \alpha_{z_1} z_{1in} + \alpha_{z_2} z_{2in} + \alpha_x x_{in} + \delta^*_{in} \tag{2}$$

$$y_{in} = 1\left[U^*_{in} = \max_{j\in C_n}\left\{U^*_{jn}\right\}\right] \tag{3}$$

This setting can be used to represent, for example, the data generation process behind a problem where an agent n chooses an alternative i among those in the choice-set C_n, following the Random Utility Maximization (RUM) model. Variables that are latent to the researcher are depicted with an asterisk and Greek letters β and α are used for the model parameters. The utility U^*_{in} and the error terms ε^*_{in} and δ^*_{in} are latent. The latter two are generated *iid* Normal $(0, \sigma^2_{\varepsilon^*})$ and Normal $(0, \sigma^2_{\delta^*})$, respectively. The researcher can only observe p_{in}, x_{in}, z_{1in}, z_{2in} and the choice y_{in}. Variables x_{in}, z_{1in} and z_{2in} are exogenous, drawn from a Normal $(0, \sigma^2_k)$, and therefore independent of the error terms of the model. Agents choose the alternative with the largest latent utility U^*_{in}, as shown in Eq. (3), where y_{in} takes value 1, if agent n chooses alternative i, and zero otherwise. Also, as shown in Eq. (1), utility U^*_{in} depends on p_{in}, x_{in} and the error term ε^*_{in}. Finally, as shown in Eq. (2), p_{in} depends on x_{in}, z_{1in}, z_{2in}, and an error term δ^*_{in}.

The structural equation, or model of interest, corresponds to Eq. (1). Endogeneity problems will arise in this equation when δ^*_{in} is correlated with ε^*_{in}, because then p_{in} will be correlated with ε^*_{in}, making it an endogenous variable. This problem may occur, for example, if ε^*_{in} contains attributes that are correlated with p_{in}, but cannot be measured by the researcher, or if there is some source of self-selection or simultaneous determination that makes the choice to be part of δ^*_{in} also. In that case, if the model shown in Eqs. (1)–(3) is estimated directly, neglecting the correlation between δ^*_{in} and ε^*_{in}, the coefficients β of the structural equation would not be retrieved consistently. Intuitively, the problem would be that, since p_{in} is correlated with ε^*_{in} through δ^*_{in}, shifts in ε^*_{in} will be erroneously endorsed to p_{in}, misinterpreting its impact in U^*_{in}, which corresponds to coefficient β_p. The inconsistency will propagate to other model coefficients, depending on the level of correlation of x_{in} with p_{in}.

Various methods may be used to correct for endogeneity in the problem depicted in Eqs. (1)–(3). The reader is referred to Guevara (2015) for an assessment of different approaches to achieve such a goal. Almost all methods rely on the availability of proper instruments, which are variables that must be relevant and exogenous. Relevance means, for the reference model depicted in Eqs. (1)–(3), that the instruments have to be correlated with the endogenous p_{in}, and exogeneity means that they have to be independent of the error term ε^*_{in} of the structural equation. It is easy to note that, for the reference model, z_{1in} and z_{2in} could play the role of instrumental variable for p_{in}.

Note also that the model described by Eqs. (1)–(3) is over-identified because it has more instruments (two) than endogenous variables (one). As it will be discussed later in Sections 3, 4 and 5, this is crucial for building the tests for the validity of instruments using overidentification.

Finally, overidentification tests will be unable to detect the endogeneity of the instruments if, as stated by De Blander (2008), the way in which they appear in Eqs. (1) and (2), are linearly dependent. For the reference model shown in

Eqs. (1)–(3) this would occur if ε^*_{in} can be written as $(\alpha_{z_1}z_{1in}+\alpha_{z_2}z_{2in})\gamma+\upsilon_{in}$, with γ being any real number and υ_{in} an exogenous error term.

The next three sections describe overidentification tests that can be used to judge instruments' exogeneity in discrete choice models. The first corresponds to the state of the art in the subject and the latter two are new tests proposed in this research. The three tests are then be assessed in Section 6 in terms of size, power, and robustness to De Blander's condition using Monte Carlo experimentation.

3. Amemiya-Lee-Newey (ALN) test

The current state of the art for the assessment of the exogeneity of instrumental variables in discrete choice models is the Amemiya-Lee-Newey (ALN) test of overidentifying restrictions. The first step in the calculation of the ALN test corresponds to the application of a two-stage estimator proposed by Amemiya (1978), which follows from replacing Eq. (2), into Eq. (1) in the reference model to obtain:

$$\begin{aligned}
U_{in} &= \beta_i+\beta_p(\alpha_0+\alpha_{z_1}z_{1in}+\alpha_{z_2}z_{2in}+\alpha_x x_{in}+\delta_{in})+\beta_x x_{in}+\varepsilon^*_{in}\\
U_{in} &= \underbrace{\beta_i+\beta_p\alpha_0}_{\pi_0}+\underbrace{\beta_p\alpha_{z_1}}_{\pi_1}z_{1in}+\underbrace{\beta_p\alpha_{z2}}_{\pi_2}z_{2in}+\underbrace{(\beta_x+\beta_p\alpha_x)}_{\pi_3}x_{in}+\underbrace{\varepsilon^*_{in}+\beta_p\delta^*_{in}}_{\tilde{\varepsilon}^*_{in}}
\end{aligned}\tag{4}$$

This equation is termed a reduced-form equation, where the right-hand side consists only of exogenous variables: instruments (z_{1in} and z_{2in}) and controls (x_{in}). Note that Eq. (2) was therefore already a reduced form equation for p_{in} in the reference model setting. By means of the transformation shown in Eq. (4), the structural model does no longer suffer from endogeneity since neither z_{1in}, z_{2in} nor x_{in} are correlated with ε^*_{in} or δ^*_{in}. Then, the estimation of this model would produce consistent estimators $\hat{\pi}$ of π_0, π_1, π_2 and π_3, up to a scale. For equivalent reasons, the estimators $\hat{\alpha}$ of α_0, α_x, α_{z_1} and α_{z_2} can be consistently estimated by directly by regressing p on x, z_1 and z_2 in Eq. (2).

It worth noting that, since ε^*_{in} and δ^*_{in} are *iid* Normal, $\tilde{\varepsilon}^*_{in}$ in Eq. (4) will also be distributed *iid* Normal, meaning that Eq. (4) could be estimated as a probit. However, as shown by Lee (1982) and Ruud (1983), consistent estimators, up to a scale, would be obtained as well when the model is estimated instead as a logit. To facilitate analysis and computation, the latter approach will be used in Section 6 for the Monte Carlo experiments developed to assess the proposed tests.

The researcher is however interested in the structural coefficients β_i, β_p and β_x, up to a scale. These can be obtained by noting that, using the estimators $\hat{\pi}$ and $\hat{\alpha}$, the following set of equations can be constructed:

$$\begin{aligned}
\hat{\pi}_0 &= \beta_i+\beta_p\hat{\alpha}_0\\
\hat{\pi}_1 &= \beta_p\hat{\alpha}_{z_1}\\
\hat{\pi}_2 &= \beta_p\hat{\alpha}_{z_2}\\
\hat{\pi}_3 &= \beta_x+\beta_p\hat{\alpha}_x
\end{aligned}\tag{5}$$

Then, the second and third expressions in Eq. (5) can be seen as observations from the following auxiliary linear model

$$\hat{\pi} = \beta_p\hat{\alpha}+w,\tag{6}$$

where w is an error term, and β_p is the only coefficient to be estimated. Therefore, given β_p, β_i, and β_x can be uniquely retrieved from the first and fourth expressions in Eq. (5).

Note that the auxiliary model in Eq. (6) makes sense only under overidentification. Otherwise, β_p would be uniquely obtainable from one of these equations. For this example, one coefficient β_p is estimable and only two observations are available because there is only one endogenous variable and two instruments. Each additional endogenous variable would result in an additional coefficient to be estimated, and each additional instrument would result in an additional observation.

To estimate this auxiliary model, Amemiya (1978) proposed a procedure, which Lee (1992) probed to the equivalent to the following minimum chi-squared or GMM estimator

$$\min_{\beta_p}\left(\hat{\pi}-\beta_p\hat{\alpha}\right)'\hat{W}^{-1}\left(\hat{\pi}-\beta_p\hat{\alpha}\right).\tag{7}$$

It can be shown (Newey, 1987) that this estimator is consistent and relatively efficient, compared to other two-stage estimators if $\hat{W}$ is a consistent estimator of the variance-covariance matrix of $(\hat{\pi}-\beta_p\hat{\alpha})$, which could be obtained as

$$\hat{W} = \widehat{\Omega}_{\Pi z}-\hat{\beta}_p\widehat{\Omega}_{\alpha z}.\tag{8}$$

$\widehat{\Omega}_{\Pi z}$ and $\widehat{\Omega}_{\alpha z}$ in Eq. (8) are, respectively, the variance-covariance matrices of the instruments estimated from the reduced form shown in Eq. (4) and from Eq. (2). $\hat{\beta}_p$ to calculate Eq. (8) can be obtained from an Ordinary Least Squares (OLS) regression of Eq. (6). Then, the inverse of $\hat{W}$ needed to calculate Eq. (7) can be obtained using a generalized inverse method (Rao and Mitra, 1971) to avoid numerical problems.

Lee (1992) noted that Amemiya's method cannot only be used to correct for endogeneity, but also to construct a test of over-identifying restrictions. The intuition behind the test is that, if the instruments are valid, the model will be consistent, and the objective function in Eq. (7) will differ from zero depending solely on the degree of overidentification of the model, solely on the number of extra instruments available per endogenous variable. In turn, if the instruments are endogenous, the estimators will be inconsistent, and the value of the objective function will be affected by the inconsistency caused by using invalid instruments.

Following Wooldridge (2010, Eq. 14.9) this GMM estimator is called a minimum chi-squared estimator because

$$S_{ALN} = \frac{1}{k_Z}(\hat{\pi} - \hat{\beta}_p\hat{\alpha})'\hat{W}^{-1}(\hat{\pi} - \hat{\beta}_p\hat{\alpha}) \sim \chi^2_{df} \tag{9}$$

follows a chi-squared distribution with degrees of freedom (*df*) equal to the degrees of overidentification of the model, which is equal to 1 in the reference model described in Eqs. (1)–(3). Therefore, the value of the objective function of Eq. (7) can be used as a test of over-identifying restrictions, appropriately standardized by sample size, which in this example is equal to the number of instrumental variables $k_Z = 2$[1].

The null hypothesis of the ALN test is that both z_1 and z_2 are exogenous instruments and the alternative hypothesis is that either z_1, z_2 or both are endogenous. This overidentification test will be consistent if De Blander's condition does not hold.

4. Refutability (REF) and modified refutability (mREF) tests

This section describes a Refutability (REF) test, which is proposed as an extension of the test for the exogeneity of instruments in linear models suggested by Card (1995), into the discrete choice framework. The REF test was originally proposed by Guevara (2010), who termed it the Direct test. Besides, it is proposed a modified version of the Refutability test (mREF) that considers all instruments simultaneously, which, as it will be shown later in Section 6, seems to show better power, size and robustness properties than its canonical counterpart.

The tests for the validity of instruments are built over a model that corrects for the endogeneity problem using overidentification. While the ALN test is built over a reduced form estimation of the model to achieve this goal, the proposed REF test is built over a control-function method correction.

The control-function method to correct for endogeneity was originally proposed by Heckman (1978). Its first applications to discrete choice modeling were developed by Rivers and Vuong (1988) and Petrin and Train (2010). For the model depicted in Eqs. (1)–(3), the control-function method can be seen as a two stage process, case in which it is termed the 2SCF. First, Eq. (2) is estimated by OLS,[2] and then the residuals of this regression are added as an auxiliary variable in the structural model of Eq. (1), which is then estimated with canonical methods. The intuition behind the 2SCF method is that, if instruments and controls are truly exogenous, the residual $\hat{\delta}_{in}$ will control for the part of the endogenous variable p_{in} that was not independent of the error term ε^*_{in} of the model. This method can also be applied by estimating both stages simultaneously in what Train (2009) termed as the Maximum Likelihood (ML) version of the control-function method.

The ML version of the control-function has the advantage of (often) increasing efficiency and facilitating the calculation of the standard errors, but at the cost of losing robustness to the underlying distributional assumption. In the Monte Carlo analysis depicted in Section 6, the ML version of the control-function method will be used. For more details about the control-function method, the reader is referred to Train (2009, Ch. 9) and Guevara (2015).

For the reference model shown in Eqs. (1)–(3), the proposed Refutability test can be described as follows. Since under the null hypothesis both instrumental variables z_1 and z_2 are truly exogenous, the model estimated using the control-function method will be consistent. Then, the inclusion of any instrument as an additional variable into the corrected choice model should produce a non-significant increase in the log-likelihood. In turn, under the alternative hypothesis that any of the instruments is endogenous, they will be correlated with the error term of the corrected model. Consequently, the inclusion of any endogenous instrument as additional variables into the model that was spuriously corrected for endogeneity with the control-function, should result in a significant increase in the log-likelihood.

Thus, the proposed REF test for the exogeneity of instruments in discrete choice models can be applied in the following three stages:

Stage 1: Estimate the reduced form equation for p_{in} (Eq. (2)) by OLS to obtain the residuals $\hat{\delta}$, as shown in Eq. (10).

$$p_{in} = \alpha_0 + \alpha_{z_1} z_{1in} + \alpha_{z_2} z_{2in} + \alpha_x x_{in} + \delta^*_{in} \Rightarrow \hat{\delta}_{in} \tag{10}$$

It can be shown (see e.g. Guevara, 2015) that if the instrumental variables z_{1in} and z_{2in} are exogenous, the error term of the model could be rewritten as $\varepsilon^*_{in} = \beta_{\hat{\delta}}\delta^*_{in} + \nu^*_{in}$, where ν_{in} will be an exogenous error. Therefore, if the estimator $\hat{\delta}_{in}$ is included in the utility, the endogeneity problem would be solved, as shown in Stage 2.

Stage 2: Estimate the structural equation (Eq. (1)) including the residuals of Eq. (10) as auxiliary variables to control for the endogeneity, and retrieve the log-likelihood of the CF estimator L^{CF}, as shown in Eq. (11).

$$U_{in} = \beta_i + \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \hat{\delta}_{in} + \tilde{\nu}^*_{in} \Rightarrow L^{CF} \tag{11}$$

[1] The extension of the test for $k_Z > 2$ follows directly from modifying Eqs. (1)–(9) to include additional z_s.

[2] Note that not only all instruments z, but also all exogenous x must be included as controls in this OLS regression explaining the endogenous variables, even if there is no a-priori reason to do it (Guevara, 2015). Only then a reduced from estimate would allow the consistent estimation of this equation.

Stages 1 and 2 can be estimated simultaneously using the ML version of the CF method, which likelihood is shown later, on Eq. (14). This will be the approach used in Section 6 for the Monte Carlo analysis.

Stage 3: Estimate the structural model again, including now not only $\hat{\delta}$, but also one of the instruments (for example z_1) as an additional variable, and retrieve the log-likelihood L^{CF_Z}, as shown in Eq. (12).

$$U_{in} = \beta_i + \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \hat{\delta}_{in} + \beta_{z_1} z_{1in} + \tilde{\tilde{v}}^*_{in} \Rightarrow L^{CF\_Z} \tag{12}$$

It should be noted that in this case is one instrument because $k_Z = 2$ in the example, but in general one could consider up to $k_Z - 1$ additional instruments in Stage 3. Also, note that stages 1 and 3 can be estimated simultaneously by the ML method using the likelihood shown in Eq. (15), which will be the approach used in Section 6 for the Monte Carlo analysis.

Finally, the statistic of the Refutability test is calculated as a likelihood ratio test in which the model estimated in Stage 2 is the restricted version of the model estimated in Stage 3, with degrees of freedom equal to degree of overidentification of the model, which in the example is equal to one.

$$S_{REF} = -2(L^{CF\_Z} - L^{CF}) \sim \chi^2_{df}. \tag{13}$$

For writing the likelihood of the problem needed to apply the REF test using the ML approach one needs to make specific distributional assumptions in Eqs. (10)–(11). For the reference model shown in Eqs. (1)–(3), δ^* is distributed *iid* Normal with variance $\sigma^2_{\delta^*}$ and, following Guevara (2015), v^* is distributed Normal with variance σ^2_{v*}. To achieve a likelihood with closed form, it can be assumed instead that v^* is distributed *iid* Extreme Value (0, $\mu_{v^*} = 1$), claiming the results of Lee (1982) and Ruud (1983). Thus, assuming also independence between observations, the likelihood of observation n can be written as shown in Eq. (14).

$$\begin{aligned} L_n^{CF} = & \frac{e^{\mu_{v+e}\left(\beta_i+\beta_p p_{in}+\beta_x x_{in}+\beta_\delta\left(p_{in}-\alpha_0-\alpha_x x_{in}-\alpha_{z_1} z_{1in}-\alpha_{z_2} z_{2in}\right)\right)}}{\sum_{j\in C_n} e^{\mu_{v+e}\left(\beta_j+\beta_p p_{jn}+\beta_x x_{jn}+\beta_\delta\left(p_{jn}-\alpha_0-\alpha_x x_{jn}-\alpha_{z_1} z_{1jn}-\alpha_{z_2} z_{2jn}\right)\right)}} \prod_{j\in C_n} \frac{1}{\sqrt{2\pi\sigma^2_{\delta*}}} \\ & \times \exp\left[-\frac{\left(p_{jn}-\alpha_0-\alpha_x x_{jn}-\alpha_{z_1} z_{1jn}-\alpha_{z_2} z_{2jn}\right)^2}{2\sigma^2_{\delta*}}\right] \end{aligned} \tag{14}$$

Eq. (14) corresponds to the ML version of the Control-Function method (see, Train, 2009, Ch 9), under a given distributional assumption, and can be used to obtain estimators of the model parameters and their standard errors. Equivalently, the ML version of the model for Stage 3 shown in Eq. (12) would have the form shown in Eq. (15).

$$\begin{aligned} L_n^{CF\_Z} = & \frac{e^{\mu_{v+e}\left(\beta_i+\beta_p p_{in}+\beta_x x_{in}+\beta_\delta\left(p_{in}-\alpha_0-\alpha_x x_{in}-\alpha_{z_1} z_{1in}-\alpha_{z_2} z_{2in}\right)+\beta_{z_1} z_{1in}\right)}}{\sum_{j\in C_n} e^{\mu_{v+e}\left(\beta_j+\beta_p p_{jn}+\beta_x x_{jn}+\beta_\delta\left(p_{jn}-\alpha_0-\alpha_x x_{jn}-\alpha_{z_1} z_{1jn}-\alpha_{z_2} z_{2jn}\right)+\beta_{z_1} z_{1jn}\right)}} \prod_{j\in C_n} \frac{1}{\sqrt{2\pi\sigma^2_{\delta*}}} \\ & \times \exp\left[-\frac{\left(p_{jn}-\alpha_0-\alpha_x x_{jn}-\alpha_{z_1} z_{1jn}-\alpha_{z_2} z_{2jn}\right)^2}{2\sigma^2_{\delta*}}\right] \end{aligned} \tag{15}$$

Just like the ALN test, the null hypothesis of the REF test is that both z_1 and z_2 are exogenous instruments and the alternative hypothesis is that either z_1, z_2 or both are endogenous. Also, this test will be consistent if De Blander's condition does not hold.

It worth noting that the null and alternative hypotheses of this REF test (and of the Direct test proposed by Guevara, 2010) differ in one key aspect from what Card (1995) considered for linear models. Card's states that one needs to assume that one instrument is exogenous, which can be used to build a corrected model over which to tests the exogeneity of other instruments. Card's approach has the limitation of not being agnostic about which instrument is exogenous or not, what can be cumbersome in practice. In this sense, the REF test seems inferior to the ALN test. However, this assumption is not really necessary. In Card's approach to the problem, the instrument that is being tested is also exogenous under the null so, for a proper correction of endogeneity, it must be included as a control in the first stage regression (see Table 5 in Card, 1995). Therefore, Card's test uses the same type of statistic considered here for the REF test, which with this re-interpretation has the advantage of considering a null that is agnostic about which instrument might be exogenous or not.

Nevertheless, despite it is true that the REF test is agnostic about which instrument is exogenous, its performance is not necessarily independent of which instrument is used to perform Stage 3, shown in Eq. (12), placing a different type of limitation on this test. In general, for the REF test only K_Z-1 out of all K_Z instruments can be included as additional variables into the third stage. This occurs because the control-function is constructed as a linear function of the endogenous variable (p) the control (x) and all instruments. Then, a model including p, x, the control-function and all instruments will be almost[3] collinear, complicating considerably the estimation of the model. This limitation is important, because it implies that one would have to explore all possible combinations of instruments to include in Stage 3, and also because it implies a loss of information, potentially reducing the power of the test.

[3] Not perfectly collinear, as stated by Guevara (2010) and apparently assumed by Card (1995).

To avoid this problem, it can be proposed a modified version of the Refutability test (mREF) that uses all the information available to reach a conclusion. As it will be shown later in Section 6, this version of the test shows, in general, larger power, smaller size distortion, and more robustness to De Blander's condition.

The mREF test begins using the same two first stages of the Refutability test, Eqs. 10 and (11), to gather the estimators $\hat{\beta}$ of Stage 2. Given this $\hat{\beta}$, the following Stage 3′ can implemented to apply the proposed mREF test.

Stage 3′: Estimate the structural model again, including $\hat{\delta}$, and considering $\hat{\beta}$ fixed. Then add all the instruments as additional variables, and retrieve the log-likelihood L^{CF_Zall}, as shown in Eq. (16).

$$U_{in} = \hat{\beta}_i + \hat{\beta}_p p_{in} + \hat{\beta}_x x_{in} + \hat{\beta}_\delta \hat{\delta}_{in} + \beta_{z_1} z_{1in} + \beta_{z_2} z_{2in} + \tilde{\tilde{\nu}}^*_{in} \Rightarrow L^{CF\_Zall} \tag{16}$$

Note that, as before, stages 1 and 3′ can be estimated simultaneously using the ML method, which will be the approach used in Section 6 for the Monte Carlo analysis.

The statistic of the mREF test is then calculated as a likelihood ratio test in which the model estimated in Stage 2 is a restricted version of the model estimated in Stage 3′ as shown in Eq. (17).

$$S_{mREF} = -2\left(L^{CF\_Zall} - L^{CF}\right) \sim \chi^2_{df}. \tag{17}$$

Equally than the REF test, the null hypothesis of the mREF test is that both z_1 and z_2 are exogenous instruments and the alternative hypothesis is that either z_1, z_2 or both are endogenous. Also, this test will be consistent if De Blander's condition does not hold and will have degrees of freedom equal to the degree of overidentification.

5. Hausman (HAU) test

An alternative method to test the validity of the instrumental variables in discrete choice model is the Hausman (HAU) test. In this case, the idea is to compare two estimators that should be consistent under the null, but one of which is more efficient than the other. The intuition behind the test is that under the null hypothesis both estimators should be similar, and under the alternative they should differ. For the linear case, Hausman (1978) derives a statistic that is Chi-square distributed and is built from the OLS and FGLS estimators and variance-covariance matrices. Hausman and McFadden (1984) does an equivalent thing, comparing a logit model estimated with a full and with a reduced choice-set, to test the validity of the Independence of Irrelevant Alternatives property. This paper builds upon Hausman and McFadden (1984) approach to test for the validity of instruments.

The proposed test consists in estimating two models that are consistent under the null hypothesis that the instruments are exogenous, but one is more efficient than the other because it uses more instruments. Just as the other tests studied, this Hausman test relies on overidentification. The first model considered is the control-function correction shown in Eq. (14), which is built using both instruments, from which the estimators $\hat{\beta}_{CF\_2}$ are obtained. The second model corresponds to the control-function estimation shown before in Eq. (18), estimated using only one instrument, from which the estimators $\hat{\beta}_{CF\_1}$ are obtained.

$$L_n^{CF\_1} = \frac{e^{\mu_{\nu+e}\left(\beta_i+\beta_p p_{in}+\beta_x x_{in}+\beta_\delta\left(p_{in}-\alpha_0-\alpha_x x_{in}-\alpha_{z_1} z_{1in}\right)\right)}}{\sum_{j\in C_n} e^{\mu_{\nu+e}\left(\beta_j+\beta_p p_{jn}+\beta_x x_{jn}+\beta_\delta\left(p_{jn}-\alpha_0-\alpha_x x_{jn}-\alpha_{z_1} z_{1jn}\right)\right)}} \prod_{j\in C_n} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left[-\frac{\left(p_{jn}-\alpha_0-\alpha_x x_{jn}-\alpha_{z_1} z_{1jn}\right)^2}{2\sigma_\delta^2}\right] \tag{18}$$

If, under the null hypothesis, both instruments z_1 and z_2 are exogenous, both estimators $\hat{\beta}_{CF\_2}$ and $\hat{\beta}_{CF\_1}$ will be consistent, and therefore similar, but $\hat{\beta}_{CF\_2}$ will be more efficient because it would make use of more information. On the other hand, if under the alternative hypothesis, z_2 is endogenous, $\hat{\beta}_{CF\_1}$ and $\hat{\beta}_{CF\_2}$ will differ. In this analysis, only the estimators of the structural model coefficients must be used in the analysis, not those of the auxiliary variables used to correct the endogeneity. Note also that, differently from ALN, REF and mREF, the HAU test is not agnostic about which instrument is valid, since one must assume that z_1 is valid to be able to estimate Eq. (18) for a proper comparison.

Following, Hausman and McFadden (1984, Eq. 1.21), the statistic of the HAU test is written as shown in Eq. (19)

$$S_{HAU} = (\hat{\beta}_{CF\_1} - \hat{\beta}_{CF\_2})'(\Sigma_{\hat{\beta}_{CF\_1}} - \Sigma_{\hat{\beta}_{CF\_2}})^{-1}(\hat{\beta}_{CF\_1} - \hat{\beta}_{C\bar{F}2}) \sim \chi^2_{df}, \tag{19}$$

where $\Sigma_{\hat{\beta}_{CF\_2}}$ is the variance-covariance matrix of the estimators of the efficient model, the one estimated using all instruments shown in Eq. (14), and $\Sigma_{\hat{\beta}_{CF\_1}}$ is the one obtained using Eq. (18). This statistic is distributed Chi-square with degrees of freedom equal to the degree of overidentification, which for the reference model shown in Eqs. (1)–(3) is equal to 1.

If worth noting that if for the estimation of the test the CF method was applied in two stages, the estimators $\Sigma_{\hat{\beta}_{CF\_2}}$ and $\hat{\Sigma}_{\hat{\beta}_{CF\_2}}$ to be used to calculate the statistic shown in Eq. (19) should not be obtained from the inverse of the Fisher Information matrix. Instead, to avoid the "estimated regressor" problem, they must either be gathered from a bootstrap estimator (see e.g. Cameron and Trivedi, 2005, Section 11) or calculated instead using the Delta method (Karaca-Mandic and Train, 2003). To avoid this problem, the Monte Carlo analysis deployed in Section 6 will be based on ML estimators of the control-function method, as shown in Eqs. (14) and (18).

Besides, since the correction for endogeneity in discrete choice models conveys in general a change of scale (Guevara and Ben-Akiva, 2012), a proper application of the proposed HAU test would only be possible if the ratios of coefficients are considered in the analysis. The easiest way to account for this, without compromising the value of the statistic, is to use a normalization in which one of the coefficients is fixed, so that the other estimators will effectively be the respective ratio. To achieve this goal, in the Monte Carlo Analysis reported in Section 6 to assess the properties of the tests under study, the price coefficient in Eq. (1) will be fixed to a value obtained for it from a preliminary estimation of the corrected model.

6. Monte Carlo experiment to assess the tests under study

6.1. Data generation process

This section reports the Monte Carlo experiment used to assess the overidentification tests for the exogeneity of instruments in discrete choice models under study. The tests are evaluated in terms of their size distortion, their power and their robustness to De Blander's condition for which all these tests are blind. The size distortion corresponds to the difference between the nominal significance of the tests, and the empirical size for the Type I error under the null. The power of the tests corresponds to the empirical probability of being able to detect cases in which the instruments are endogenous under various circumstances. These first two measures are the usual tools for the assessment of statistical test (see, e.g. Cohen; 1988). The last measure is specific for the type of overidentification tests considered in this study. It corresponds to the analysis of potential differences in the degree of the inability of the proposed tests to detect endogeneity under De Blander's condition, and on its vicinity.

For expositional purposes the experiments are built as a special case of the reference model described in Eqs. (1)–(3). The qualitative conclusions attained are then validated with a sensitivity analysis regarding sample size, degrees of freedom, degree of endogeneity of the instruments and the distributional assumption for the error terms.

The data generation process consists of a binary probit choice model in which the utility of each alternative depends linearly on its price *p*, a control *x*, and an error term ε^*, which is divided into two components, ξ^* and e^*. ξ^* represents an omitted attribute that is correlated with *p*, and e^* is an *iid* error distributed Normal (0,1). As a first sensitivity analysis, not reported here for the sake of space, the experiments were fully regenerated considering that e^* was instead *iid* Extreme Value (0,1), reaching the same qualitative results, with only negligible numerical differences.

The model coefficients used for the data generation process are shown in Eq. (20). The sample size is $N=2000$ observations for all the cases analyzed, except when studying size, where *N* was varied to analyze its impact.

$$U_{in} = -1p_{in} + 1x_{in} + \underbrace{2\xi^*_{in} + e^*_{in}}_{\varepsilon^*_{in}} \tag{20}$$

The price *p* was constructed as a function of ξ^*, the control *x*, and two exogenous variables z_1 and z_2, as shown in Eq. (21)

$$p_{in} = 2\xi^*_{in} + 0.5z_{1in} + 0.5z_{2in} + 0.5x_{in} + \delta^*_{in}, \tag{21}$$

where δ^*, ξ^*, *x*, z_1 and z_2 are generated *iid* Normal (0,1). Under this setting, if ξ^* is omitted in the specification of the utility in Eq. (20), *p* will be correlated with the error term ε^* causing endogeneity. In addition, z_1 and z_2 will be valid instruments because they are correlated with *p* and independent of ε^*.

To analyze the size, power and robustness properties of the tests, the data generation process considers also the following two variables

$$\begin{aligned} b_{1in} &= \lambda_1\xi^*_{in} + (1-\lambda_1)z_{1in} + \psi^*_{1in} \\ b_{2in} &= \lambda_2\xi^*_{in} + (1-\lambda_2)z_{2in} + \psi^*_{2in}, \end{aligned} \tag{22}$$

where λ_1,$\lambda_2 \in [0, 1]$ and ψ^*_{1in} and ψ^*_{2in} were generated *iid* Normal (0,1).

Variables b_1 and b_2 will be used as candidate instrumental variables, to assess the ability of the tests under study to detect their eventual endogeneity, which will be defined by the value of a given combination of $\langle\lambda_1,\lambda_2\rangle$. b_1 will be an exogenous instrument when $\lambda_1=0$ and it will be endogenous for any value $\lambda_1 > 0$, with a degree that would grow with λ_1 up to $\lambda_1=1$. The same will occur for b_2 with λ_2.

To analyze the impact of the degrees of freedom, the data generation process was extended to include in some cases a third instrumental variable z_3 in Eq. (21), which was generated also *iid* Normal (0,1) and with $\alpha_{z_3}=0.5$. Likewise, to achieve this goal, a candidate instrumental variable b_3 was generated as shown in Eq. (22), considering a respective λ_3.

The study was performed by implementing 2000 repetitions for each setting of the data generation process, and reporting the percentage of times each test under study rejects the null hypothesis with a nominal value of 5% for the Type I error. Despite the data was generated using normal errors, to avoid the prohibitive computational cost that would have been involved in the estimation of tens of thousands of probit models, the estimation was performed using logit models, which have a closed form that preclude the need for integration and, following the results of Lee (1982) and Ruud (1983), is a suitable approximation for the problem under study.

The precision attained in the simulation experiments was calculated as the standard error of the count of the rejections among 20 sets of 100 realizations of the data generation process considered in each case. The 2000 repetitions allowed achieving a 1% precision, or more, in the estimation of the rate of rejection for all cases. The significant digits are reported accordingly.

Table 1
Assessment of size distortion. Empirical % of Type I error at nominal 5%.

Test	ALN		REF		mREF		HAU	
N	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$
150	4	1	9	4	2	3	10	6
300	9	2	12	5	4	4	11	6
500	10	4	11	5	3	2	13	4
1000	15	7	11	4	3	3	19	4
2000	19	10	11	5	3	3	19	8
5000	20	11	11	4	3	3	19	7

$\lambda_1=\lambda_2=\lambda_3=0$; 2000 repetitions. % rate of rejections with 1% precision.

6.2. Assessment of size distortion

To analyze the size distortion of the tests, the analysis is performed by a simulation that considers a case in which $\lambda_1=\lambda_2=\lambda_3=0$, so that all b_s are valid instrumental variables. A size distortion will exist when, under this setting, the simulated rate of rejection of the null hypothesis that the instruments are valid differs from the nominal 5% imposed. Empirical rejections below the nominal 5% are said to be “conservative” and values above are said to be “liberal”. The former is interpreted as something desirable since it means that the test results in fewer wrong rejections of the null hypothesis than expected. The latter is an undesirable property of the test for the contrary reason.

The size is analyzed as a function of sample size N for the Amemiya-Lee-Newey (ALN), the Refutability (REF), the modified Refutability (mREF) and the Hausman (HAU) tests. Results are also assessed in terms of the degrees of freedom, considering a case when $k_z=2$ and $k_z=3$ instrumental variables. Results are reported in Table 1.

The first column of Table 1 contains the results attained for the ALN test, which is the current state of the art for judging the exogeneity of instrumental variables in nonlinear models. Results show that ALN has a size that grows with the sample size and only seems to stabilize for very large samples. Besides, for $N=5000$ the test shows a large liberal size distortion when the test is estimated with only one degree of freedom ($k_z=2$), and the problem is reduced, but does not vanish, for $k_z=3$.

The REF test shows better results in the second column of Table 1. First of all, the size seems to stabilize for samples as small as 500. Besides, the REF test exhibits a large liberal size distortion for $k_z=2$, but the problem seems to be solved for $k_z=3$, when the nominal and empirical size are numerically the same, up to the precision attained.

The mREF, in the third column of Table 1, shows even better results. First of all, just as the REF test, the size of the mREF test seems to stabilize for samples as small as 500. However, the mREF test also shows a small conservative size distortion (something that is desirable) for almost all sample sizes and degrees of freedom analyzed. It can be hypothesized that the improvement, compared to the REF test, may be the result of using all the information available with the same sample size.

Finally, the HAU test shows inferior results, in the last column of Table 1. The size seems to stabilize for $N=1000$ and there is a larger liberal size distortion which, despite it decreases with the degrees of freedom, does not seem to vanish. Besides, the estimation of the HAU test proved to be more unstable, since for 3% of the simulations for $N=150$ and $k_z=2$, the estimation failed, possibly because of the need for re-scaling and for estimating the inverse of a matrix.

6.3. Assessment of power

To assess the power of the tests, the analysis is performed via a simulation that considers a case in which $\lambda_1=\lambda_3=0$, so that b_1 and b_3 are valid instrument, and λ_2 varies from 0 to 1 in 0.1 steps. The study is performed reporting the rate of rejection attained with each test as the degree of endogeneity grows with λ_2. The tests with larger rejection rate for a given degree of endogeneity are said to have more power and should be preferred. Results are reported in Table 2.

The first column of Table 2 summarizes the results obtained for the Amemiya-Lee-Newey (ALN) test. When $\lambda_2=0$ the null hypothesis that the instruments are exogenous is true, and the rate of rejection corresponds to the size of the test. The results are concordant with what was reported in Table 1 for $N=2000$, considering the 1% precision attained. Regarding the power of the ALN test, as expected, as λ_2 grows from 0.1 to 1, increasing the degree of endogeneity of b_2, it becomes easier for the ALN test to detect the problem, already reaching a 100% rejection of the null hypothesis when $\lambda_2=0.4$. Comparing the case with one and two degrees of freedom, for $k_z=2$ there is a large liberal size distortion, which is reduced with $k_z=3$, and the shape of the power curve is adjusted accordingly.

The second, third and fourth column of Table 2 summarize the results attained for three versions of the Refutability (REF) test. REF (b_1), reported in column 2, corresponds to the test described in Eq. (13) calculated using b_1 as the instrument that is being tested (included in Eq. 12), which is always exogenous because $\lambda_1=0$. REF (b_2), reported in column 3 of Table 2, is build using instead b_2 as the instrument being tested, which endogeneity grows with λ_2. The third one, mREF, reported in column 4 of Table 2, corresponds to the test described in Eq. (17) calculated using all available instruments simultaneously.

First, it can be remarked that the results for REF (b_1) and REF (b_2) are numerically the same, up to the precision level attained, when $k_z=2$. This means that, with one degree of freedom, the REF test is not only agnostic about the exogeneity

Table 2
Analysis of power. Empirical % of rejection at nominal 5%.

Method	ALN		REF (b_1)		REF (b_2)		mREF		HAU (b_1)		HAU (b_2)	
λ_2	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$
0	18	10	11	4	11	4	2	3	19	7	20	7
0.1	46	30	64	24	64	63	52	53	83	71	0	1
0.2	87	77	98	71	98	99	97	99	94	91	0	0
0.3	98	98	100	96	100	100	100	100	98	99	0	0
0.4	100	100	100	100	100	100	100	100	100	100	16	2
0.5	100	100	100	100	100	100	100	100	100	100	79	40
0.6	100	100	100	100	100	100	100	100	100	100	98	93
0.7	100	100	100	100	100	100	100	100	100	100	100	97
0.8	100	100	100	100	100	100	100	100	99	100	93	70
0.9	100	100	100	100	100	100	100	100	99	99	61	26
1	100	100	100	100	100	100	100	100	100	99	32	6

$\lambda_1=\lambda_3=0$, $N=2000$, 2000 repetitions, % rate of rejections with 1% precision.

of the instruments, but also independent of the sequence in which the instruments are used to build the test. However, with more degrees of freedom the results for REF *(b_1)* and REF *(b_2)* do differ. When $k_z=3$ REF *(b_2)* shows larger power than REF *(b_1)*, possibly because b_2 is the endogenous instrument in this experiment, what facilitates the detection of the endogeneity problem. Regarding the mREF test, as also shown in Table 2, it shows larger power as the worst case of the REF test, which occurs when using the exogenous instrument b_1. Summarizing, mREF test shows results that are superior to those of the REF test, not only because it is indifferent to the quality of the instruments that are being tested, but also because it achieves larger power when it is properly compared. Furthermore, both the REF and the mREF tests studied show larger power than the state of the art ALN test to assess the exogeneity of instruments.

The last two columns of Table 2 summarize the results attained for the Hausman (HAU) test proposed in this article. Two versions of the test are reported. HAU (b_1), reported in column 5, corresponds to the HAU test in which b_1 is used to build the estimator that is obtained with a single instrument, as shown in Eq. (18). This is the proper application of the HAU test in the sense that, in this experiment, b_1 is always an exogenous instrument, something that, however, the researcher will never now in practice. HAU (b_2) corresponds to the HAU test applied using instead the endogenous instrument b_2 to build the estimator that is obtained with a single instrument. Results show that only HAU (b_1) has the expected behavior of a power that grows with λ_2. In turn, HAU (b_2) does not only shows a low power, but also an erratic behavior with a power that sometimes goes up and sometimes goes down. This occurs because, under this setting, b_2 is endogenous, precluding the proper application of the HAU test. This is problematic for the application of the HAU test in practice, because the researcher never knows which instrument might be exogenous, making it then difficult to reach a strong conclusion from the test. Regarding the comparison of the HAU test with other tests, HAU (b_2) is clearly inferior and, although HAU (b_1) shows slightly larger power for some particular cases (e.g. $\lambda_2=0.1$ compared with mREF) the difference seems barely circumstantial and seemingly be the result of an undesirable liberal size distortion.

6.4. Assessment of the robustness to De Blander's condition

The third property studied is the robustness of the tests to the very peculiar condition for which De Blander shows that overidentification tests are inconsistent. As it was explained in Section 2, De Blander's condition under this experimental setting, for k_z=2, corresponds to a case in which b_{1in} and b_{2in} appear in Eq. (21) as $\alpha_{b1}b_{1in}+\alpha_{b2}b_{2in}$ and also in ε^*_{in}, in the structural equation of the utility in Eq. 20, in the form $(\alpha_{b_1}b_{1in}+\alpha_{b_2}b_{2in})\gamma$, with γ being any real number. An equivalent condition can be built for the case of $k_z=3$.

To study the robustness of the tests for the exogeneity of instruments to De Blander's condition, the instrumental variables under study will again be generated as shown in Eq. (22), but now considering that $\lambda_1=\lambda_3=0.5$. Under this setting, for all values of λ_2 at least one instrument will always be endogenous, something that should be detected by the tests under study. However, the tests will be unable achieve their purpose when De Blander's condition hold, something that, under this setting, will occur when $\lambda_2=0.5$, since then $\text{cov}(b_1,\xi*)=\text{cov}(b_2,\xi*)=\text{cov}(b_3,\xi*)$ and, at the same time, $\text{cov}(b_1,p)=\text{cov}(b_2,p)=\text{cov}(b_3,p)$, fulfilling the required condition.

Although all three overidentification tests under study are blind to De Blander's condition, it is relevant to identify potential differences in size, when this condition holds exactly, and the difference in power when the data generation processes is on its vicinity. To analyze this, the value of λ_2 will be varied from 0 to 1 in 0.1 steps and in Table 3 it will be reported the percentual rate, among the 2000 repetitions, that each test rejects the null hypothesis with a nominal value of 5% for the Type I error.

The first column of Table 3 reports the results attained for the ALN test. It can be noted that when $\lambda_2=0.5$, when De Blander's condition holds exactly, ALN test shows empirical power equal to zero. The test was not able to detect that the instruments were endogenous on almost any of the 2000 simulation repetitions. For values of $\lambda_2\neq 0.5$ ALN shows power

Table 3
Analysis De Blander's Hypothesis. Empirical % of rejection at nominal 5%.

Method	ALN		REF (b_1)		REF (b_2)		mREF		HAU (b_1)		HAU (b_2)	
λ_2	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$	$k_z=2$	$k_z=3$
0	100	96	100	49	100	100	100	100	79	97	100	26
0.1	99	77	100	47	100	100	100	100	82	96	100	23
0.2	83	37	100	42	100	98	100	99	86	94	98	19
0.3	38	6	88	27	88	81	88	87	87	82	82	13
0.4	5	0	46	15	46	34	46	48	71	55	43	10
0.5	0	0	17	9	17	8	17	21	34	16	34	16
0.6	4	0	39	16	39	33	39	46	5	2	53	26
0.7	22	4	76	36	76	78	75	83	0	0	67	37
0.8	55	24	95	63	95	97	94	98	0	0	79	48
0.9	82	62	99	84	99	100	99	100	0	0	86	58
1	95	88	100	95	100	100	100	100	0	0	91	63

$\lambda_1=\lambda_3=0.5$, $N=2000$, 2000 repetitions, % rate of rejections with 1% precision.

that grows slightly asymmetrically as getting further away from De Blander's condition. Besides, results in Table 3 show that the rate of growth of the power of the ALN is smaller with more degrees of freedom.

Second, third and fourth column of Table 3 report the results attained for three versions of the REF test. For one degree of freedom, REF (b_1), REF (b_2) and mREF show numerically the same results, up to the reported 1% significance, which is coherent with what was reported in Table 2 with large amounts of endogeneity. The power under De Blander's condition for $\lambda_2=0.5$ is far from zero (17%) and it grows symmetrically up to 100% for $\lambda_2 \neq 0.5$. The situation changes with two degrees of freedom, $k_z=3$. The size of both REF (b_1) and REF (b_2) goes down at $\lambda_2=0.5$, and the behavior of REF (b_1) outside De Blander's condition becomes clearly worse, never reaching a 100% power, not even for $\lambda_2=0$. The reason behind this finding may be that, under this experimental setting, the instrument that is being tested in that case (b_1) always remains close to De Blander's condition, independent of the value of λ_2. On the contrary, the robustness of mREF test improves with more degrees of freedom, showing a larger power for all values of λ_2.

Finally, consider the HAU test. Although the size for HAU (b_1) and HAU (b_2) for $\lambda_2=0$ and $k_z=2$ is larger than that of the mREF test, the behavior of the HAU test seems rather erratic and sometimes inferior for other values of λ_2, precluding it from being the recommendable tool for practice. Noticeably, HAU (b_1) shows zero power for large values of λ_2. This may be attributable to the fact that, contradicting the assumption used to build the statistic, the model with more instruments (e.g b_1 and b_2) results in larger variance than the model estimated only with b_1, because the endogeneity of b_2 becomes too large.

Summarizing, under this specific experimental setting, the state of the art Amemiya-Lee-Newey (ALN) test is clearly surpassed in terms of robustness to De Blander's condition by Refutability (REF) test and particularly by its modified version (mREF). Besides, the Hausman (HAU) test showed a rather erratic behavior, discouraging its use. Given these results on robustness, and previous ones on power and size shown in Sections 6.2 and 6.3, a general recommendation for practice could be to use the mREF test.

7. Conclusion

Endogeneity is a pervasive problem in all econometric models, including discrete choice models, and has been habitually neglected in transportation science and practice. Methods to correct for endogeneity rely on the availability of instrumental variables that must comply with the competing properties of being correlated with the endogenous variable but, at the same time, exogenous to the problem. Verifying the second condition is especially difficult because the error term of the model is not observed. This article extends the tool-box of statistical test to achieve this goal in discrete choice models in general. Besides, the proposed tests are compared to the state of art in terms of size, power and their robustness to the condition for which all overidentification tests are blind.

The state of the art for testing the exogeneity of instruments in discrete choice models is the Amemiya-Lee-Newey (ALN) test, which relies in the estimation of an auxiliary GMM model build from reduced-form estimates. The two alternative tests proposed in this paper are constructed as adaptations of the Refutability (REF) test and the Hausman (HAU) test, into the discrete choice framework. The REF test consists in including instruments as additional variables in an auxiliary model that was corrected for endogeneity using the control-function method, considering all instruments available. The HAU test is built from the comparison of the estimates attained using different subsets of instruments.

Variations of the three tests were assessed using a binary choice Monte Carlo experiment. Results suggests that the modified Refutability (mREF) test, which includes all instruments simultaneously, should be recommended, as it has larger power, smaller size distortion, and more robustness, compared other test analyzed. The HAU test has the limitation of not being agnostic about which instrument might be exogenous, and the REF has the limitation of potentially needing to explore all possible combinations of instruments to conclude. These results were validated by means of a sensitivity analysis regarding sample size, degrees of freedom, degree of endogeneity of the instruments and the distributional assumption of the error term. Despite there are no obvious alternative hypotheses for which these findings may be qualitatively different, it should be reminded however that these conclusions are circumscribed to the case studies analyzed, and any extension to other experimental settings should be carefully assessed.

Finally, regarding future lines of research on this topic, it would be interesting to explore the power and size of the tests for the exogeneity of instruments under other circumstances. Some alternatives for this could be to use diverse choice model settings, to consider different nominal values for the Type I error, to use probit for estimation, more degrees of freedom, and to use real data gathered from different sources.

Acknowledgments

This publication was funded in part by CONICYT, FONDECYT 1150590 and by the Complex Engineering Systems Institute (CONICYT: FB0816). Preliminary work for this research was developed as part of the Doctoral dissertation of Guevara (2010, Ch.4). All Monte Carlo experiments were generated, estimated and analyzed using the open-source software R (R Development Core Team, 2008).

References

Amemiya, T., 1978. The estimation of a simultaneous equation generalized probit model. Econometrica 46, 1193–1205.
Bresnahan, T.F., 1997. Comment. In: Bresnahan, T.F., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, Chicago.
Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: methods and applications. Cambridge university press.
Card, D., 1995. Using geographic variation in the college proximity to estimate the return to schooling. In: Christophides, Garnt, Swidinsky (Eds.), Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp. University of Toronto Press, Toronto, Canada, pp. 201–222.
Cohen, J., 1988. Statistical Power Analysis For the Behavioral Sciences. Lawrence Erlbaum Associates, Publishers, New York.
De Blander, R., 2008. Which null hypothesis do overidentification restrictions actually test? Econ. Bull. 3 (76), 1–9.
Fernández-Antolín, A., Guevara, C.A., de Lapparent, M., Bierlaire, M., 2016. Correcting for endogeneity due to omitted attitudes: empirical assessment of a modified MIS method using RP mode choice data. J. Choice Modell. 20, 1–15.
Ferreira, F., 2010. You can take it with you: proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities. J. Public Econ. 94 (9-10), 661–673.
Guevara, C.A., 2010. Endogeneity and Sampling of Alternatives in Spatial Choice Models Ph.D. Thesis. Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
Guevara, C.A., 2005. Addressing Endogeneity in Residential Location Models Master Thesis. Massachusetts Institute of Technology.
Guevara, C.A., 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transp. Res. Part A 82, 240–254.
Guevara, C.A., Ben-Akiva, M.E., 2012. Change of scale and forecasting with the control-function method in logit models. Transp. Sci. 46 (3), 425–437.
Guevara, C.A., Polanco, D., 2016. Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution. Transportmetrica A 12 (5), 458–478.
Guevara, C.A., Thomas, A., 2007. Multiple classification analysis in trip production models. Transp. Policy 14 (6), 514–522.
Guevara, C.A., Tang, Y., Gao, S., 2017. The initial condition problem with complete history dependency in learning models for travel choices. Transp. Res. Part B doi:10.1016/j.trb.2017.09.006, Fothcomming.
Guevara, C., Ben-Akiva, M., 2006. Endogeneity in residential location choice models. Transp. Res. Rec. 1977, 60–66.
Guevara, C.A., Tirachini, A., Hurtubia, R. and Dekker, T. (2018) Correcting For Endogeneity Due to Omitted Crowding in Public Transport Choice Using the Multiple Indicator Solution (MIS) Method. Working paper, Universidad de Chile.
Guevara, C.A. and Navarro, P. (2015) Detection of weak instruments when correcting for endogeneity in binary logit models. IATBR 2015-WINDSOR.
Hausman, J., 1978. Specification tests in econometrics. Econometrica 46, 1251–1272.
Hausman, J., 1997. Valuation of new goods under perfect and imperfect competition. In: Bresnahan, T.F., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, Chicago.
Hausman, J., McFadden, D., 1984. Specification tests for the multinomial logit model. Econometrica 1219–1240.
Heckman, J., 1978. Dummy endogenous variables in a simultaneous equation system. Econometrica 46, 931–959.
Helveston, J.P., 2016. Development and Adoption of Plug-in Electric Vehicles in China: Markets, Policy, and Innovation PhD Thesis. Carnegie Mellon Univeristy.
Karaca-Mandic, P., Train, K., 2003. Standard error correction in two-stage estimation with nested samples. Econometrics J. 6 (2), 401–407.
Lee, L., 1982. Specification error in multinomial logit models. J. Econometrics 20, 197–209.
Lee, L., 1992. Amemiya's generalized least squares and tests of overidentification in simultaneous equation models with qualitative or limited dependent variables. Econometric Rev. 11 (3), 319–328.
Lurkin, V., Garrow, L.A., Higgins, M.J., Newman, J.P., Schyns, M., 2017. Accounting for price endogeneity in airline itinerary choice models: an application to Continental US markets. Transp. Res. Part A 100, 228–246.
Mumbower, S., Garrow, L.A., Higgins, M.J., 2014. Estimating flight-level price elasticities using online airline data: a first step toward integrating pricing, demand, and revenue optimization. Transp. Res. Part A 66, 196–212.
Newey, W., 1985. Generalized method of moments specification testing. J. Econometrics 29, 229–256.
Newey, W., 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. J. Econometrics 36, 231–250.
Nichols, A., 2007. Causal inference with observational data. Stata J. 7 (4), 507–541.
Parente, P.M.D.C., Silva, J.M.C.S., 2012. A cautionary note on tests of overidentifying restrictions. Econ. Lett. 115, 314–317.
Petrin, A., Train, K., 2010. A control function approach to endogeneity in consumer choice models. J. Marketing Res. 47 (1), 3–13.
Pleus, M., 2015. Implementations of Tests on the Exogeneity of Selected Variables and Their Performance in Practice PhD Thesis. Tinbergen Institute, Amsterdam.
R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org.
Rao, C., Mitra, S., 1971. Generalized Inverse of a Matrix and Its Applications. J. Wiley, New York, NY.
Rivers, D., Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. J. Econometrics 39 (3), 347–366.
Ruud, P., 1983. Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete models. Econometrica 51, 225–228.
Sargan, J., 1958. The estimation of economic relationships using instrumental variables. Econometrica 26, 393–415.
Stock, J., 2001. Instrumental variables in statistics and econometrics. In: Smelser, Baltes (Eds.), International Encyclopaedia of the Behavioural Sciences. Elsevier Publishing, New York, pp. 7577–7582.

Train, K.E, 2009. Discrete choice methods with simulation. Cambridge university press.

Wen, C.H., Chen, P.H., 2017. Passenger booking timing for low-cost airlines: a continuous logit approach. J. Air Transp. Manage. 64, 91–99.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT press.

Zegras, P.C., Li, M., Kilic, T., Lozano-Gracia, N., Ghorpade, A., Tiberti, M., Aguilera, A., Zhao, F., 2018. Assessing the representativeness of a smartphone-based household travel survey in Dar es Salaam, Tanzania. Transportation 1–29.