Article:

Guevara, CA, Tirachini, A, Hurtubia, R et al. (1 more author) (2020) Correcting for endogeneity due to omitted crowding in public transport choice using the Multiple Indicator Solution (MIS) method. Transportation Research Part A: Policy and Practice, 137. pp. 472-484. ISSN 0965-8564

https://doi.org/10.1016/j.tra.2018.10.030

Correcting for Endogeneity due to Omitted Crowding in Public Transport Choice Using the Multiple Indicator Solution (MIS) Method

C. Angelo Guevara[1*], Alejandro Tirachini[1], Ricardo Hurtubia[2] and Thijs Dekker[3]

1 Transport Engineering Division, Civil Engineering Department, Universidad de Chile, Santiago, Chile.
2 School of Architecture and Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Santiago, Chile.
3 Institute for Transport Studies (ITS), University of Leeds, UK.
* Corresponding author: crguevar@ing.uchile.cl

Accepted for Publication in Transport Research Part A – October 2018

ABSTRACT

Crowding levels are very relevant for the analysis and evaluation of the performance of public transport as they strongly affect the level of service and the overall perceived quality of the system. However, crowding is not an easy variable to measure and, hence, demand models often tend to ignore or use abstract proxies for it. In this paper we assess the Multiple Indicator Solution (MIS) method in a Stated Preference (SP) experiment where crowding conditions were displayed to the respondent, but are artificially omitted in the estimation to cause endogeneity. Results show strong evidence that the MIS method can be used to control for a wide range of omitted attributes in SP data. We also discuss the potential application of this approach to Revealed Preferences (RP) models of public transport by asking suitable post-trip questions to users. Two MIS variations were applied to this SP case study and both provided outcomes that were superior to those of the curtailed model. We enrich the analysis with the aid of Monte Carlo simulation. Results suggest that potential problems may arise in the presence of neglected interactions and if indicators are only weakly correlated with the omitted attribute. For the SP case study analysed, only the former issue seems to play a role in the results. The article finishes by discussing the implications of these findings for the correction of endogeneity on SP and RP data on public transport.

Keywords: Stated Preference (SP), Stated Choice (SC), Crowding, Comfort, Security, Proxy

1. INTRODUCTION

Passengers' crowding is a key component of the perceived level of service of public transport systems (Cox et al., 2006; Wardman and Whelan, 2011; Mohd Mahudin et al., 2012). A few countries (e.g., France, Sweden, Australia and the United Kingdom) have included the effect of crowding discomfort in their official guidelines for transport project appraisal (OECD/ITF, 2014). In these countries the discomfort of crowding has been translated into crowding penalties (or multipliers) over the baseline Value of Travel Time (VTT). The crowding multipliers were recently reassessed in the UK as part of a new national VTT study commissioned by the UK Department for Transport (Batley et al., 2017). Beyond effects on users' comfort, crowding has a direct impact on various aspects of public transport supply, including optimal vehicle frequency, size and fare; as well as on other components of the level of service, such as travel time variability, waiting, and in-vehicle times (Tirachini et al., 2013).

Measuring the level of crowding and its impact of (non-)public transport users is not trivial. The measurement of objective proxies of crowding in the field is complicated because crowding occurs under high-occupancy levels, where measurement may be extremely problematic. Moreover, crowding is a diffuse concept, associated with the subjective interpretation of the physical phenomenon represented by a high density of persons sharing a limited space (Cox *et al*., 2006; Mohd Mahudin *et al.,* 2012). Although it has often been characterized with measures of density in terms of standees per square metre (Whelan and Crockett, 2009, Tirachini *et al*., 2013), load factor (ratio of number of passengers to number of seats, see Wardman and Whelan, 2011) or as the probability of getting a seat (Hensher et al., 2011), what really matters to passengers includes a fuzzy mix of aspects related to safety, security, privacy, physical comfort, smell, freedom of movement and many other variables.

In most public transport systems there are no data available directly measuring or allowing the indirect estimation of occupancy levels of vehicles, because boarding, and specially alighting, are not ubiquitously recorded. Even when boarding and alighting of individual passenger is recorded, measurement might still be problematic when weekly or monthly passes exist and their holders are not required to validate their passes. When aggregated data of boardings and alightings is available (for example, smartcard data or with vehicles equipped with automatic passenger counting devices), it is possible to estimate the load profile of a rail or metro line to obtain aggregate occupancy measures. Newer technologies like mobile phone data have also been used to estimate train occupancy (Aguiléra *et al*., 2014). However, these figures will at most be average measures for a whole line or a whole train, not considering that passengers may distribute themselves unevenly in carriages, or that consecutive trains may have different occupancies. Thus, even with advanced passenger data available, it is difficult to measure the real occupancy level, let alone the perceived level of crowding that is faced by public transport users. In this context, including the level of perceived security or comfort in behavioural models is a novel approach to internalise the effect of crowding exposure.

Omission or imperfect measurement of the perceived level of crowding, may make a demand model unsuitable for policy analysis, especially when assessing a transport project that reduce peak or average crowding levels. If crowding is ignored, the demand model will most likely suffer from endogeneity, an econometric problem that arises when the error term of the model is not independent of the model variables. Endogeneity may be the result of the omission of attributes, in this case crowding, errors in variables or simultaneous determination. In this context, traditional estimation methods will fail to retrieve consistent estimators of the model parameters, resulting, e.g., in wrong policy recommendations and demand predictions (see, e.g Guevara and Thomas, 2007). This calls for new estimation tools (see, e.g. Guevara, 2015). In this article we illustrate and assess the applicability of the Multiple Indicator Solution (MIS) method (Guevara and Polanco, 2016) to address the problem of endogeneity due to omitted crowding in public transportation choice models.

Since perceived crowding is likely to be correlated with travel time, its omission from the model will likely cause endogeneity. As a result, the impact of travel time on route or mode choice will be overestimated (see Tirachini *et al.*, 2013) since shocks in crowding and travel time may be confounded. A model built from Revealed Preference (RP) or Stated Preference (SP) data omitting crowding, may overestimate the value of travel time savings and would be blind to policies targeted at alleviating crowding levels, like changing from single to double-deck or articulated buses. Such model will also underestimate the value of increasing the frequency of service in public transport systems (Jara-Díaz and Gschwender, 2003) or of increasing the seat-capacity of a line (Tirachini *et al.*, 2014; 2016).

The proposed MIS method relies on having at least a pair of suitable indicators for crowding. The indicators are measured variables that depend on the omitted variable causing endogeneity and may be collected in RP or SP experiments, e.g., by asking the interviewees to recall a previous experience, collecting data from social networks or even by actively measuring, e.g. the load of the vehicle. The relative easiness in collecting such data makes the MIS an attractive tool for the correction of endogeneity in public transportation choice models.

The MIS method is applied in two stages. First, one of the indicators is included as an explanatory variable in the utility of the choice model. By this modification, the endogeneity of other variables is eliminated, and the included indicator becomes the only endogenous variable. Then, in a second stage, the problem is solved by using the second indicator as an instrument for the first one. Wooldridge (2010) applies the MIS method to linear models; where he proposes using the 2SLS method for the second stage. Guevara and Polanco (2016) adapted and applied the MIS method to Logit models by considering the control-function approach (Heckman, 1978; Petrin and Train, 2010; Guevara and Ben-Akiva 2006, 2012; Guevara, 2015; Fernández et. al, 2016) in the second stage of the MIS method.

From a qualitative point of view the MIS seems to have several advantages for the problem at hand over alternative approaches such as latent variable models. Guevara (2015) shows that while the latent variable approach may be more efficient, its application requires stronger distributional assumptions and the availability of proper structural equations for the latent variable. Both assumptions are not needed for the MIS method and would be difficult to gather in an RP application.

The quantitative assessment of the MIS method is performed by re-analysing the data from an ad-hoc SP survey in which we artificially cause endogeneity and have proper proxies and indicators to address the problem. The full details of the SP experiment are described in Tirachini et al. (2017). In the experiment the interviewees were asked to choose between two alternatives. Each choice required the interviewees to make a trade-off between crowding levels, standing and travel time in Metro (subway) trips in Santiago, Chile. The level of crowding was illustrated to the interviewee by a picture, a diagram (see Figure 1 in Section 2), or a text description. All visual representations were designed to meet a pre-specified objective level of crowding density. For example, Table 1 depicts the text descriptions and the respective percentage of standees and occupied seats used to build the profile for each crowding level considered. Additionally, a block of the interviewees was asked afterwards to qualify the illustrations alone, in terms of “comfort” and “security (against pickpocketing and physical and verbal abuse)” in general, using a Likert scale.

We artificially introduce endogeneity in this experiment by estimating a curtailed model in which we omit the information about the level of crowding. The MIS method is then applied to correct for endogeneity using the reported levels of comfort and security as indicators. In this way, we mimic a field experiment in which crowding conditions were not properly measured, but perception of crowding was collected by a post-trip questionnaire applied to the traveller or to third users. The success of the MIS method is then assessed comparing it with a model that includes the objective level of crowding density in the model, which in this SP case study, works as a perfect proxy, but would hardly be available in RP.

Two MIS versions are explored: MIS Comfort and MIS Security. The former considers the inclusion of comfort in the utility and uses security as the instrument for the application of the control-function method. The latter reverses the role of the indicators. We also enrich the analysis with the aid of Monte Carlo simulation to explore the potential impact of the failure of some modelling assumptions.

Table 1: Representation of Crowding Level by Text Description

Crowding level	Text Description (shown to respondents)	Description (not shown to respondents)
1	Less than half or seats are occupied. No one is standing.	35% seats occupied, 0 standees
2	More than half or seats are occupied. No one is standing.	69% seats occupied, 0 standees
3	All seats are occupied. Few people standing, there is no difficulty moving.	100% seats occupied, 1 pax/m2 standing
4	All seats are occupied. People standing, minor difficulty moving.	100% seats occupied, 2 pax/m2 standing
5	All seats are occupied. Many people standing, it is difficult to move.	100% seats occupied, 4 pax/m2 standing
6	All seats are occupied. Maximum number of people standing, maximum difficulty to move.	100% seats occupied, 6 pax/m2 standing

The article is structured as follows. After this introduction, Section 2 briefly describes the SP data available. Then, Section 3 describes the MIS method and qualitatively discusses the applicability of it on this SP case study. Section 4 conveys the details of the application of the MIS method and the analysis of the results attained. Section 5 reports the study, using Monte Carlo simulation, of some potential limitations of the method. The article finishes discussing the implications of these findings for the correction of endogeneity in RP data on public transport.

2. DESCRIPTION OF THE CASE STUDY

The main survey used in the case study is part of a research project on the valuation of crowding discomfort in the Metro system in Santiago, Chile. The survey (for details see Tirachini et al. 2017) collects respondent characteristics (gender, age, income, occupation, car ownership and access to car, smartphone availability/use during metro trips, crowding perception and time use) alongside an SP experiment consisting of six binary choice tasks asking each respondent to choose between two metro trips varying in crowding levels amongst other attributes.

Three attributes were used to characterize the alternatives: travel time, crowding level, and if the passenger has to travel sitting or standing. Travel time in the SP survey is pivoted around the knowledge base of travellers, that is, the reported travel time on his/her latest metro trip. Five attribute levels were set for travel time (-25%, -12.5%, 0, +12.5% and

+25% of latest travel time). It should be noted that the null variation in travel time of this pivoted SP experiment does not correspond to the complete "recent trip" because crowding and being or not seated are varied. Accordingly, the experiment does not suffer of the type of endogeneity described by Train and Wilson (2008). In total, a design comprising twelve different choice tasks were generated, grouped into two blocks of six tasks of which only one was presented to each respondent.

The crowding attribute has six levels and was presented in three different types of illustrations across alternative versions of the survey: namely with text, 2D diagrams (bird´s-eye view), and photos taken inside a metro car. With reference to Table 1, the crowding levels range from zero (almost empty train) to 6 (completely full train). In Figure 1 we depict an example of a choice task as shown to respondents, in which the train occupancy level is depicted with 2D diagrams. In Alternative 1, the passenger is standing in a crowded train (4 pax/m2) and travel time is 19 minutes, whilst in Alternative 2 the passenger is sitting in a less crowded train (0 pax/m2) and travel time is 31 minutes.

Choose your preferred alternative:

Alternative 1	Alternative 2
- Travel time: **19 minutes**	- Travel time: **31 minutes**
- You travel **standing** under these conditions:	- You travel **sitting** under these conditions:

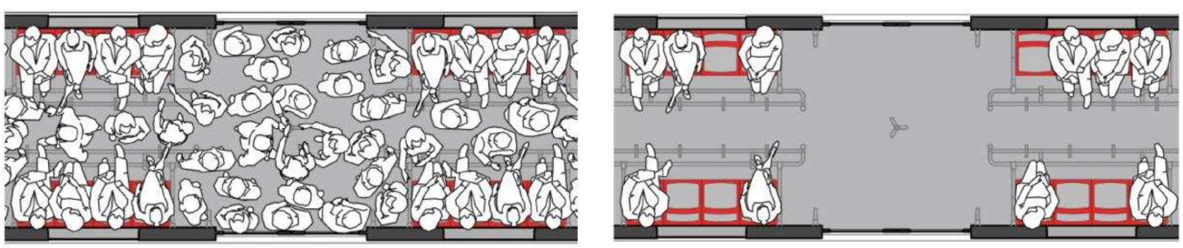

Figure 1: Example of SP task in which crowding is illustrated by a 2D diagram

The survey was programmed on the online survey platform Qualtrics. A pilot study (n=25) was carried out in September 2014, and the final survey was conducted in October 2014 by a private consultant. In the pilot, the SP survey was designed using an orthogonal design; whereas for the final survey a D-efficient design (Rose *et al.*, 2009) was generated using the SP experimental design software NGene.

Two survey-platforms were used: (a) online, in which the survey was distributed by email to a panel of respondents from the consultant, (b) face-to-face, in which surveyors with tablets interviewed metro users outside selected stations. The total number of correct complete surveys is 413 (210 online surveys, 203 face-to-face surveys). The sampling strategy attempted to resemble the income profile of Santiago´s metro users, as shown in Tirachini et al. (2017).

3. ON THE APPLICABILITY OF THE MIS METHOD TO OMITTED CROWDING DISCOMFORT

We follow the Random Utility Maximisation (RUM) model and assume that an individual *n* will select either the left or right option presented to him (or her) depending on which one generates the highest level of indirect utility.

As shown in Eq. (1), indirect utility U_{in}^{*} is assumed to be linear, with coefficients β, on three attributes: travel time (*tt*); whether the individual is standing or not (*st*); and an illustration of the crowding conditions (*cr*). The utility is complemented by an exogenous additive error term e_{in}^{*}.

$$U_{in}^{*} = \beta_{i} + \beta_{tt} tt_{in} + \beta_{st} st_{in} + \beta_{cr} cr_{in}^{*} + e_{in}^{*} \qquad (1)$$

It may be possible that passengers' sensitivity to crowding (β_{cr}) may be larger for longer trips or for trips performed standing. To account for such an effect, researchers usually interact crowding related attributes (e.g. density of passengers) with travel time or other variables when modelling passengers' sensitivity to crowding (see reviews by Wardman and Whelan 2011, Tirachini *et al.*, 2013). Since in this case study we only have indicators for crowding provided outside the SP experiment (i.e. not linked to other attributes presented in the choice task), for applying the MIS method, we must assume that such an effect is negligible. This does not preclude, of course, the existence of correlation between (tt_{in}) and/or standing $(st_{in})$with *cr*, something that will almost always occur and is the source of the endogeneity problem. Our only assumption is that β_{cr} does not vary along these dimensions.

The asterisk (*) in Eq. (1) and elsewhere is used to depict the variables that are latent to the researcher that wants to estimate the model. $cr*$ is latent because it is unclear for the researcher how to account for what the interviewee inferred from the illustration of the crowding level. The utility U_{in}^{*} is also latent, but the researcher observes the choice that, under the random utility maximization (RUM) framework, corresponds to variable y_{in}, as shown in Eq. (2), which takes value 1 if alternative i has the largest random utility.

$$y_{in} = 1\left[U^*_{in} \geq U^*_{jn} \mid \forall j \in C_n\right] \tag{2}$$

Since cr^* is latent and it is ignored in the model estimation, the actual error term of the choice model (Eq. (1)- (2)) will be $\varepsilon^*_{in} = \beta_{cr} cr^*_{in} + e^*_{in}$ and the utility will take the form shown in Eq. (3).

$$U^*_{in} = \beta_i + \beta_{tt} tt_{in} + \beta_{st} st_{in} + \varepsilon^*_{in} \tag{3}$$

Estimating the model shown in Eq. (3) causes endogeneity because the experimental design forced trade-offs between crowding levels and tt and st, respectively (see, for example, Figure 1). Therefore, both variables will be negatively correlated with *cr*. Consequently, a model that omits cr^* in this case study will confound the effect of large crowding levels with, e.g., short tt and st, resulting in estimated coefficients for both variables that are more positive than what they really are in the population. The model shown in Eq. (3) will be called Endogenous (curtailed) Model later in the experiments reported in Table 2.

Because of the way in which this SP experiment was designed, it is possible to account for the endogeneity problem using the Proxy method, something that, in turn, would unlikely be feasible for RP data. The key thing is that the illustrations for the level of crowding deployed in each experiment were made based on pre-specified levels of density *dd* (see Table 1), which can then be used as perfect proxies for *cr* , in an econometric sense (see Guevara, 2015).

Formally, *dd* can be a perfect proxy for *cr*, because in the design of the experiment the researcher followed a sequence of actions that allows assuming a chain of causality between the density *dd* and the perceived level of crowding *cr*. Indeed, the researcher first defined *dd*, then prepared an illustration of it, and then asked the choice maker to choose from a profile that contained that illustration. Consequently, the level of crowding *cr* experienced by an individual can be described by a function of the pre-specified level of density *dd* and an exogenous error term γ^* , as shown in Eq. (4).

$$cr^*_{in} = \theta_0 + \theta_d dd_{in} + \gamma^*_{in} \tag{4}$$

Then, if Eq. (4) is substituted into Eq. (1) we obtain expression (5)

$$U^*_{in} = \beta_i + \beta_{tt} tt_{in} + \beta_{st} st_{in} + \beta_{cr} \theta_d dd_{in} + \tilde{\varepsilon}^*_{in}, \tag{5}$$

where the error term $\tilde{\varepsilon}_{\text{in}}^{*}$ is

$$\tilde{\varepsilon}_{\text{in}}^{*} = \beta_{\text{cr}}\left(\theta_0 + \gamma_{\text{in}}^{*}\right) + \text{e}_{\text{in}}^{*}, \tag{6}$$

which is orthogonal to all model variables *tt*, *st* and *dd*, solving the endogeneity problem. This implies that all model coefficients ($\beta_{\text{i}}, \beta_{\text{tt}}, \beta_{\text{st}}, \beta_{\text{cr}}\theta_{\text{d}}$) will be consistently estimated, up to a scale, if *dd* is introduced in the model as a proxy for the omitted crowding. Utility function (5) is called Proxy (Benchmark) model, as estimated in Table 2.

It should be noted that although *dd* is a perfect proxy for *cr*, it neglects the fact that the perception of crowding is related with individual characteristics (Cox *et al.*, 2006). However, such potential limitations may be addressed, for example, considering that the coefficient θ_d is heterogeneous across agents using latent classes, a random coefficient, or a systematic taste variation approach.

As shown by Guevara and Ben-Akiva (2012), the scale (variance) of the model corrected for endogeneity is lower (higher) than the scale (variance) of the original model. This is a direct result of working with $\tilde{\varepsilon}_{in}^{*}$ instead of e_{in}^{*}. As a result, only the ratio of parameter estimates is meaningful to compare across models. We will concentrate our analysis on the ratio β_{st}/β_{tt} , which represents the rate of substitution between being standing and travel time.

Note that, from Eq. (4), it follows that only the product $\beta_{cr}\theta_d$ can be identified, but not β_{cr} itself, not even up to a scale. This implies that the model insights on the behavioural responses to a change in the level of crowding, will depend on the quality of the proxy used. Similar identification issues arise with any other feasible method correcting for endogeneity using proxies or indicators (Guevara, 2015).

The proxy approach described above can only be applied here because of the way in which the data was generated in this SP experiment. This is useful because it allows us to compare the MIS method against a benchmark model. A similar benchmark model cannot be constructed in any RP experiment in which the indicator of crowding is obtained indirectly, for example, through:

I1: Measuring the weight of train cars or the buses;

I2: Measuring the loads indirectly by processing passive data gathered from smartcards or cell-phones;

I3: Asking post-trip questions to the passengers about their perception of the level of crowding experiences, or

I4: By processing images taken on-board the vehicles.

The problem that arises in the above cases is that, instead of the actual density, an imperfect measure can be obtainable. For example, if a noisy version *xd* of *dd* is collected, such that

$$\mathrm{xd}_{\mathrm{in}} = \mathrm{dd}_{\mathrm{in}} + \varphi^{*}_{\mathrm{in}}, \tag{7}$$

where φ^{*} is an exogenous error term, the model resulting from using *xd* instead of *dd* will be affected by additional endogeneity due to measurement error. Indeed, substituting the above equation into Eq. (5) gives:

$$\begin{aligned} \mathrm{U}^{*}_{\mathrm{in}} &= \beta_{\mathrm{i}} + \beta_{\mathrm{tt}}\mathrm{tt}_{\mathrm{in}} + \beta_{\mathrm{st}}\mathrm{st}_{\mathrm{in}} + \beta_{\mathrm{cr}}\theta_{\mathrm{d}}\left(\mathrm{xd}_{\mathrm{in}} - \varphi_{\mathrm{in}}\right) + \tilde{\varepsilon}^{*}_{\mathrm{in}} \\ \mathrm{U}^{*}_{\mathrm{in}} &= \beta_{\mathrm{i}} + \beta_{\mathrm{tt}}\mathrm{tt}_{\mathrm{in}} + \beta_{\mathrm{st}}\mathrm{st}_{\mathrm{in}} + \beta_{\mathrm{cr}}\theta_{\mathrm{d}}\,\mathrm{xd}_{\mathrm{in}} - \underbrace{\beta_{\mathrm{cr}}\theta_{\mathrm{d}}\varphi_{\mathrm{in}} + \tilde{\varepsilon}^{*}_{\mathrm{in}}}_{\tilde{\tilde{\varepsilon}}^{*}_{\mathrm{in}}} \end{aligned}, \tag{8}$$

where the correlation between φ^{*} and *xd* introduce endogeneity by construction.

In turn, a feasible method to correct for endogeneity if an error-free measure of density is not available in the RP context, is the Multiple Indicator Solution (MIS) (Guevara and Polanco 2016). The key component of the method is the availability of at least two indicators - of the types I1, I2, I3, I4 described above - of the latent variable that causes the endogeneity. In the context of our SP survey, the pair of indicators is obtained from the responses provided to crowding perception questions. More specifically, a subset of the interviewees was presented with an illustration of a level of crowding, not necessarily the one experienced during the trip on which the interviewees were intercepted (for the data collected in the field), neither those illustrated in the choice profiles presented to them. For this crowding level the interviewee was asked the following two questions:

1) “from 1 ‘very insecure’ to 7 ‘very secure’, how secure do you feel travelling under these conditions?”;
2) “from 1 ‘very uncomfortable’ to 7 ‘very comfortable’, how comfortable do you feel travelling under these conditions?”.

Three out of the six possible crowding levels (see Table 1) were presented to each respondent, depicted either using the diagrams, pictures or text descriptions. These crowding levels were presented alone, without additional information on travel time or seat use.

The responses to these questions, *ss* (security) and *cc* (comfort) respectively, work as indicators for the level of crowding *cr* inferred by the respondents from the illustrations (picture, diagram or text), in a similar way that post-trip questions would work in RP data. The key is that the way in which the indicators were collected imply a chain of causality that allows assuming that the following equations hold:

$$\mathrm{ss}_{\mathrm{in}} = \eta_{\mathrm{ss}}^{0} + \eta_{\mathrm{ss}}^{\mathrm{cr}} \mathrm{cr}_{\mathrm{in}}^{*} + \mathrm{e}_{\mathrm{ss\_in}}^{*} \qquad (9)$$

$$\mathrm{cc}_{\mathrm{in}} = \eta_{\mathrm{cc}}^{0} + \eta_{\mathrm{cc}}^{\mathrm{cr}} \mathrm{cr}_{\mathrm{in}}^{*} + \mathrm{e}_{\mathrm{cc\_in}}^{*}, \qquad (10)$$

where the η's are model parameters and the e^{*}'s are exogenous error terms assumed to be independent not only to *cr*, but also to *tt* and *st*.

If one of the indicators, e.g. *ss*, is substituted into the structural equation of utility (i.e. Eq. (1)) below, then the model will still suffer of endogeneity because $e_{cc\_in}^{*}$ in $\tilde{\tilde{\varepsilon}}_{in}$ is correlated with cc_{in} by construction, as shown in Eq. (9).

$$\mathrm{U}_{\mathrm{in}}^{*} = \beta_{\mathrm{i}} + \beta_{\mathrm{tt}} \mathrm{tt}_{\mathrm{in}} + \beta_{\mathrm{st}} \mathrm{st}_{\mathrm{in}} + \frac{\beta_{\mathrm{cr}}}{\eta_{\mathrm{cc}}^{\mathrm{cr}}} \left(\mathrm{cc}_{\mathrm{in}} - \eta_{\mathrm{cc}}^{0} - \mathrm{e}_{\mathrm{cc\_in}}^{*} \right) + \mathrm{e}_{\mathrm{in}}^{*}$$

$$\mathrm{U}_{\mathrm{in}}^{*} = \beta_{\mathrm{i}} + \beta_{\mathrm{tt}} \mathrm{tt}_{\mathrm{in}} + \beta_{\mathrm{st}} \mathrm{st}_{\mathrm{in}} + \frac{\beta_{\mathrm{cr}}}{\eta_{\mathrm{cc}}^{\mathrm{cr}}} \mathrm{cc}_{\mathrm{in}} \underbrace{- \frac{\beta_{\mathrm{cr}}}{\eta_{\mathrm{cc}}^{\mathrm{cr}}} \left(\eta_{\mathrm{cc}}^{0} + \mathrm{e}_{\mathrm{cc\_in}}^{*} \right) + \mathrm{e}_{\mathrm{in}}^{*}}_{\tilde{\tilde{\varepsilon}}_{\mathrm{in}}} \qquad (11)$$

However, by this transformation, *cc* became the only endogenous attribute in this modified model. Furthermore, this endogeneity can be addressed with the control-function method because ss is a valid instrument for the endogenous variable cc. Indeed, ss is correlated with cc because both depend on *cr* through Eq. (8) and Eq.(9) but, at the same time, ss will be independent of $\tilde{\tilde{\varepsilon}}_{\mathrm{in}}$ as long as it is independent from $\mathrm{e}_{\mathrm{cc\_in}}^{*}$.

Under these assumptions consistent estimators of the model parameters, up to a scale, can be obtained by applying the following two-stage procedure:

Stage 1: Regress the endogenous variable *cc* on the controls *tt*, *st* and the instrument *ss* to obtain the residuals $\hat{\delta}$

$$\mathrm{cc}_{\mathrm{in}} = \alpha_0 + \alpha_1 \mathrm{tt}_{\mathrm{in}} + \alpha_2 \mathrm{st}_{\mathrm{in}} + \alpha_3 \mathrm{ss}_{\mathrm{in}} + \delta_{\mathrm{in}} \quad \Rightarrow \hat{\delta}_{\mathrm{in}}. \qquad (12)$$

In this auxiliary regression the variables are stacked by alternative and individuals and the residuals $\hat{\delta}$ calculated in this way capture all the part of the endogenous cc that was correlated with the error term of the model (see. e.g Guevara, 2015).

Stage 2: Estimate the choice model considering *tt*, *st*, *cc* and $\hat{\delta}_{cc}$ in the utility function (13). This is the MIS Comfort model in Table 2.

$$\mathrm{U}_{\mathrm{in}}^{*} = \beta_{\mathrm{i}} + \beta_{\mathrm{tt}}\mathrm{tt}_{\mathrm{in}} + \beta_{\mathrm{st}}\mathrm{st}_{\mathrm{in}} + \frac{\beta_{\mathrm{cr}}}{\eta_{\mathrm{cc}}^{\mathrm{cr}}}\mathrm{cc}_{\mathrm{in}} + \beta_{\delta}\hat{\delta}_{\mathrm{in}} + \mathrm{v}_{\mathrm{in}}^{*}, \tag{13}$$

to obtain consistent estimators $\hat{\beta}$ of $\beta_{\mathrm{i}}, \beta_{\mathrm{tt}}, \beta_{\mathrm{st}}, \frac{\beta_{\mathrm{cr}}}{\eta_{\mathrm{cc}}^{\mathrm{cr}}}$ up to a scale. Note that, like the perfect proxy model described above, only $\frac{\beta_{\mathrm{cr}}}{\eta_{\mathrm{cc}}^{\mathrm{cr}}}$ and not β_{cr} can be identified in the corrected model.

Besides, the role of *cc* and *ss* can be reversed attaining, asymptotically, the same result. Finite sample properties may, however, be considerably different. In Section 4, the model in which the role of cc and ss are reversed is called MIS Security. Analogously to the MIS Comfort model, the MIS Security model is defined by the following equations for stages 1 and 2.

$$\mathrm{ss}_{\mathrm{in}} = \tilde{\alpha}_0 + \tilde{\alpha}_1\mathrm{tt}_{\mathrm{in}} + \tilde{\alpha}_2\mathrm{st}_{\mathrm{in}} + \tilde{\alpha}_3\mathrm{cc}_{\mathrm{in}} + \tilde{\delta}_{\mathrm{in}} \quad \Rightarrow \hat{\tilde{\delta}}_{\mathrm{in}}. \tag{14}$$

$$\mathrm{U}_{\mathrm{in}}^{*} = \beta_{\mathrm{i}} + \beta_{\mathrm{tt}}\mathrm{tt}_{\mathrm{in}} + \beta_{\mathrm{st}}\mathrm{st}_{\mathrm{in}} + \frac{\beta_{\mathrm{cr}}}{\eta_{\mathrm{ss}}^{\mathrm{cr}}}\mathrm{cc}_{\mathrm{in}} + \beta_{\tilde{\delta}}\hat{\tilde{\delta}}_{\mathrm{in}} + \tilde{\mathrm{v}}_{\mathrm{in}}^{*}, \tag{15}$$

Note also that If more than two indicators are available, the same procedure can be applied by including all additional indicators in the first stage OLS regression.

The MIS can be estimated sequentially or simultaneously by Full Information Maximum Likelihood (FIML). The sequential application is much easier to implement, but it may compromise efficiency and requires the estimation of the standard errors via bootstrap or the method proposed by Karaca-Mandic and Train (2003). The sequential estimation approach will, however, be more robust to alternative modelling assumptions.

The model assumed that the β_{cr} parameter is constant across the population, which may be a questionable assumption since it has been claimed that crowding is a personal perception of a given situation (Mohd Mahudin *et al.* 2012). This effect could be captured in Eq. (1) if the coefficient β_{cr} can vary across the population in the form of a random coefficient, latent classes or systematic taste variations. The Proxy and the MIS methods could both be applied under such circumstances. For the latter, a different set of indicators would be needed for each class.

Finally, the indicators used for the application of the MIS in this SP case study cannot be directly the ones given by each respondent. The problem is that, as it was explained before, only a subset of the interviewees was presented with an illustration of a level of crowding that was not necessarily the same deployed in their choice profiles and neither for all alternatives. To address this limitation, the indicators used for the MIS in the case study were built as the average responses given by all other interviewees that were questioned about the given level of crowding deployed in a profile.

4. APPLICATION OF THE MIS METHOD TO CORRECT FOR ENDOGENEITY DUE TO OMITTED CROWDING

This section reports the estimation of four models with the SP data on public transport choice under the omission of crowding information. Reading from left to right in Table 2, the curtailed or endogenous model (Eq. 3) presents biased parameter estimates. This is a direct result from omitting the crowding variable without applying a correction. Because the survey design considered a trade-off between crowding and other model variables, the omission of crowding produced a large positive finite bias for β_{tt} and β_{st}. The bias was large enough to make both estimators positive, although β_{tt} is not significantly different from zero. The estimator β_0 is positive and significant, suggesting some response bias towards choosing the alternative at the right.

Table 2: Estimation Results. Correcting for Endogeneity with Different Methods

	Endogenous (curtailed)			Proxy (Benchmark)			MIS Comfort			MIS Security		
	$\hat{\beta}$	s.e	p	$\hat{\beta}$	s.e	p	$\hat{\beta}$	s.e.	p	$\hat{\beta}$	s.e.	p
β_0	0.147	0.0404	0%	0.159	0.0433	0%	0.170	0.0415	0%	0.170	0.0431	0%
β_{tt}	0.00358	0.00534	50%	-0.114	0.00928	0%	-0.0993	0.00867	0%	-0.0977	0.00980	0%
β_{st}	0.187	0.0672	1%	-0.245	0.0759	0%	-0.0202	0.0754	79%	-0.157	0.0824	6%
$\tilde{\beta}_{cr}$				-0.339	0.0199	0%	0.396	0.0227	0%	0.466	0.0295	0%
β_{δ}							-0.749	0.0846	0%	0.447	0.111	0%
β_{st} / β_{tt}	52.3	70.0	46%	2.14	0.566	0%	0.204	0.752	79%	1.60	0.766	4%
N	2467			2467			2467			2467		
$\bar{\rho}^2$	0.00466			0.0871			0.0949			0.0949		
$L(0)$	-1709.99			-1709.99			-1709.99			-1709.99		
$L(\hat{\beta})$	-1699.03			-1557.04			-1542.69			-1542.69		

s.e: Standard errors, calculated by bootstrap for Proxy and MIS and with the delta method for β_{st}/β_{tt} ; p: p-value for $\beta = 0$ test . β_0 : ASC for the alternative of the right. β_{tt} : Travel time; $\tilde{\beta}_{cr}$: Density for Proxy and indicator included in utility for MIS; $\tilde{\beta}_{cr}$ = $\beta_{cr}\theta_d$ for Proxy model, $\frac{\beta_{cr}}{\eta_{cc}^{cr}}$ for MIS Comfort and $\frac{\beta_{cr}}{\eta_{ss}^{cr}}$ for MIS Security; β_{st} : 1 if the person was standing and zero otherwise. β_{δ} : residual of the auxiliary regression of the MIS method, $\beta_{\tilde{\delta}}$ for MIS security.

The second model in Table 2 (Eq. 5) acts as our benchmark model and corresponds to the estimators obtained using the density (standees per square meter) for the perceived level of crowding. As discussed in section 3, density is a perfect proxy for crowding. Note that the same variable has been applied in previous studies by e.g., Whelan and Crockett (2009) and Tirachini *et al.*, (2013, 2016). Now, the estimators of β_{tt} and β_{st} both have the expected sign and are significantly different from zero. The same applies to β_{cr}. The value of the $\bar{\rho}^2$ increased 20 times, compared to the endogenous model, showing that the level of crowding depicted was an important attribute in the choice process.

The ratio of the estimators $\beta_{\mathrm{st}}/\beta_{\mathrm{tt}}$ is also presented in Table 2 and shows a large difference between the curtailed (endogenous) and proxy model. The shift is also observed in Figure 2, where the 95% confidence interval for the ratio of the Endogenous model is even out of the figure.[1]

The third model (MIS Comfort) corresponds to the application of the MIS method including Comfort in the choice model and Security as the instrument, as described in equations (11), (12) and (13). Security (ss) is used to regress comfort (cc), as shown in Eq. (12), and the residual of this regression is then used to estimate the utility shown in Eq. (13). This model is substantially better than the endogenous one because the coefficients of β_{tt} and β_{st} now have a negative sign and their ratio is much closer to the value obtained with the Proxy model (Table 2 and Figure 2). The fit of the MIS Comfort is slightly better than the one obtained for the proxy model.[2] This suggests that, for this specific case, the combination of the two indicators provided more information than the *dd* proxy alone. There are, however, two shortcomings of the MIS Comfort model. First, the Hausman and McFadden´s (1984) test rejects the null hypothesis that β_{st} / β_{tt} is the same in the Proxy model and the MIS Comfort model (p-value 30%). This is a problem since the proxy model acts as the benchmark model that does not suffer of endogeneity. Second, the p-values of the estimator of β_{st} and the ratio β_{st} / β_{tt} are unacceptably low. Both limitations are solved for the final model.

Finally, the fourth model (MIS Security) corresponds to the application of the MIS method using Security in the utility and Comfort as the instrument, as shown in Eqs. (14) and (15). Again, the model fit improves over the Endogenous model and is comparable to that of the MIS Comfort model. In contrast, the Hausman-McFadden test does not allow us

[1] To formally assess this difference, one may use a variation of the Hausman and McFadden (1984) test. However, the test proved to be uninformative in this case, because of the large variance of β_{ST} in the endogenous model.

[2] The Horowitz (1983) test for non-nested hypotheses shows that the probability that the proxy model is better, even though the rho-squared of the MIS Comfort is larger, is below 4%.

to reject the null-hypothesis of equality of the ratio β_{st}/β_{tt} between the MIS Security and the benchmark model (p-value 0%). Hence, the MIS Security model properly corrected for endogeneity. These results are further detailed in Figure 2, where the MIS Security confidence interval of the estimated β_{st}/β_{tt} contains almost completely the confidence interval of the Proxy model. As expected, the confidence interval is slightly wider (less efficient) due to the use of indicators but no longer suffers from endogeneity. In the MIS Security model, the estimators of β_{st} and β_{st}/β_{tt} are significant, with p-values of 6% and 4% respectively. Hence, this model specification does not suffer from the issues encountered in the MIS Comfort model.

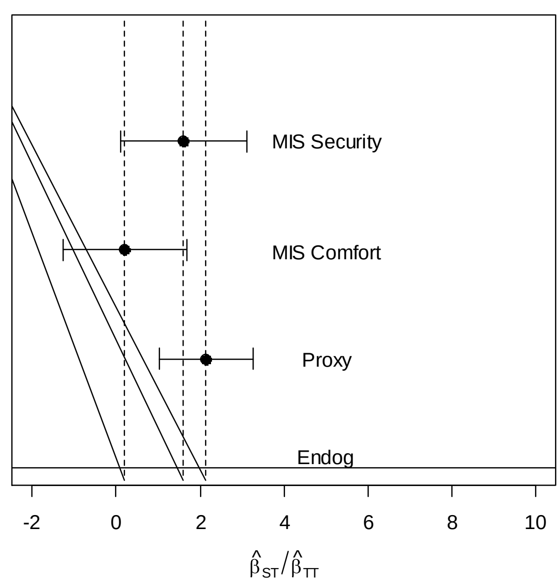

Figure 2: 95% Confidence Intervals for β_{st}/β_{tt} with Various Estimation Methods

Summarizing, these results show strong evidence that both MIS versions provided outcomes that were superior to those of the endogenous model, but only the one including Security in the utility produced estimators that were matching the ones attained in the benchmark model using a perfect proxy. The MIS application showed better fit, even after correcting by degrees of freedom, suggesting that the combination of the two indicators provided more information than the *dd* proxy acting alone. A question that remains is why the version of the MIS method that used Comfort as an instrument (MIS Security), provided much better results that its reversed counterpart.

One possibility could be the use of weak instruments in the second stage of the MIS method (see also Guevara 2015). We explore this issue further in Section 5. However, this argument could be discarded. The F-test of the first stage of the control function are 5.25e+04 for MIS Comfort and 4.94e+04 for MIS Security, both four orders of magnitudes larger than the thresholds that have been suggested for linear models (see, e.g. Stock and Yogo 2002), which were preliminary suggested to hold for discrete choice models by Guevara and Navarro (2015).

Another possibility for the difference found could be a failure of the redundancy assumption. That is, either tt or st are also at the right-hand side of Eqs (8) and (9) explain in part the realization of the indicators. Usually, this cannot be tested in practice but, because we have *dd,* which acts as a perfect proxy for *cr*, we can explore this issue by means of OLS regressions. In these regressions the indicators are explained by the values of *dd*, *tt* and *st* as presented in the choice experiment. Since the choice model is binary, the number of observations of this auxiliary linear regression is twice the number of observations of the choice model. If the redundancy assumption holds, the coefficients of *tt* and *st* should be equal to zero. The results of the experiment to test the redundancy assumption are presented in Table 3.

It can be noted in Table 3 that *tt* is significant for the comfort indicator and *st* is significant for the security indicator. A comparison with the ratios of the respective estimator with that of *dd* in the respective model, it can be noted that the potential failure is two orders of magnitude larger for the Security indicator ($\alpha_{\mathrm{st}}/\alpha_{\mathrm{dd}} = -0.231$), than for the Comfort indicator ($\alpha_{\mathrm{tt}}/\alpha_{\mathrm{dd}} = -0.0019$). This implies that the potential violation of the redundancy assumption seems to be mild for the Comfort indicator, but not negligible for the Security one.

Table 3: Analysis of the Redundancy Assumption

	Comfort Indicator			Security Indicator		
	$\hat{\alpha}$	s.e.	p-value	$\hat{\alpha}$	s.e.	p-value
α_0	7.01	0.0153	0%	6.56	0.0188	0%
α_{dd}	-1.10	0.00364	0%	-0.851	0.00447	0%
α_{tt}	0.00131	0.000419	0%	-0.000764	0.000515	14%
α_{st}	0.0240	0.0130	7%	0.197	0.0160	0%
$\bar{R}^2$	0.9581			0.8963		
N	4934			4934		

As shown by Guevara and Polanco (2016), the MIS method can also provide good results if only one of the indicators fails the redundancy assumption, which seems to be the case in this application. Problem seem to arise when the poorer Security indicator is used as an instrument in the MIS Comfort application. This may explain why the MIS Security model performs better.

Three possible hypotheses could be suggested as potential explanations for the failure of the redundancy assumption. The first hypothesis is that, because the indicators were collected after the SP experiment, the respondents implicitly considered the levels of other SP attributes (*st* and *tt*), despite they were explicitly instructed not to do so. However, this hypothesis can immediately be discarded because this potential issue was controlled for by eliminating the indicator provided by the incumbent individual from the respective average indicator. The second hypothesis is related to model misspecification specification (i.e. the linearity of Eqs. (8) and (9)). Usually, this cannot be tested but, again, since we have *dd* as a perfect proxy for *cr*, we may explore the validity of the linearity assumption.

Figure 3 depicts the boxplot and the average value of the indicators as a function of the crowding level (see Table 1). It can be noted that the relation of the comfort indicator is almost linear for all but the level 1 (35% seats occupied, 0 standees), which has virtually the same declared comfort as level 2 (69% seats occupied, 0 standees). Instead, in the case of the Security indicator, the relation is much more concave across the range of crowding levels. The Pearson´s correlation coefficient for Comfort is -0.979 and for Security is -0.945. This result suggests that the failure of the linearity assumption could be another explanation of the poor performance of Security as an instrument for Comfort in the MIS Comfort model.

The potential problem with the neglected nonlinearity of the security indicator is that the size of its error will depend on the value of *cr*. Moreover, *cr* is nonlinearly related with *tt* and *st* through the SP design. Thus, the security indicator will correlate with both *tt* and *st* not only through *cr*, breaking the redundancy assumption. In Section 5 we develop a Monte Carlo Experiment in which we explore the potential impact of the failure of the linearity assumption. Results suggest that, even if the linearity assumption fails significantly, it may only have a negligible impact in the ability of the MIS method to correct for the endogeneity.

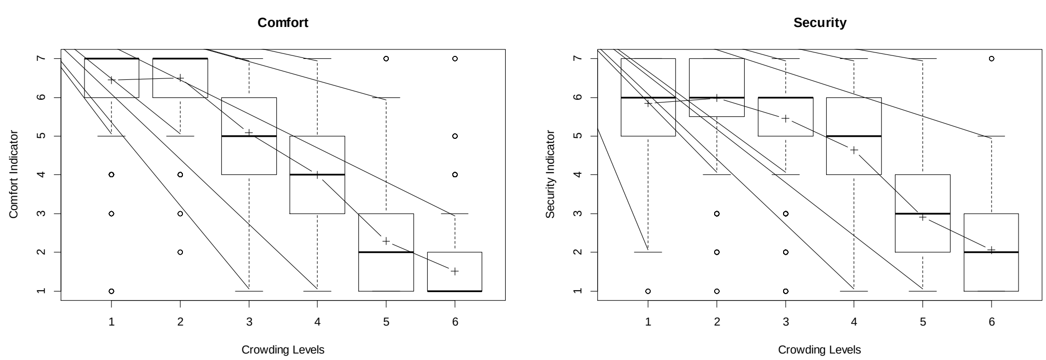

Figure 3: Boxplot and Average (+) of Indicators as a Function of Crowding Level

The third hypothesis for the failure of the redundancy assumption is that the impact of the level of crowding in the utility does not depend only on the density depicted in the illustration, but also on the other model attributes depicted in the choice task, i.e. travel time or whether the trip was done standing or not. This would be a problem because the indicators were gathered directly from the illustration, after the choice task, without showing other model attributes. Fernández-Antolin et al (2016) have shown that the problem of misspecification of the utility function through neglected interaction effects can be addressed if available indicators account directly for the interaction. However, this only holds if the indicator is gathered from a non-interacted omitted attribute. In a real RP experiment, this assumption may hold, but in SP experiments this is not the case, precluding the redundancy assumption to hold~~.~~

To study this third hypothesis, we re-estimated the Proxy model reported in Table 2 by adding first the product between density and travel time and then between density and standing. Results show strong evidence that the interaction with travel time is not statistically significant (p-value 17%). However, the hypothesis that the interaction of density and standing is equal to zero is clearly rejected (p-value 0%). This suggests the hypothesis that neglected interaction may be playing a role our case study in the failure of the redundancy assumption is true. Besides, since the security indicator correlates with standing in Table 3 this explains why MIS Comfort does not work as opposed to MIS Security.

Summarizing, among the three possible hypotheses that could potentially explain the failure of the redundancy assumption for MIS Comfort only possible neglected interaction seems to be behind this finding. The potential impact of this and other modelling issues will be analysed in the next section using Monte Carlo analysis.

5. MONTE CARLO ANALYSIS OF THE ROBUSTNESS OF THE MIS METHOD

In this section we use Monte Carlo simulation to study changes in the performance of the MIS method to variations in sample size, the degree of linearity of the indicator with respect to the omitted attribute, the strength of the indicators, the degree of endogeneity and the impact of neglected interaction. By this, we enrich the analysis of the case study deployed in Section 4 and complement the work on the MIS method developed before by Guevara and Polanco (2016) and Fernandez-Antolin et al (2016). The study of the impact of the degree of linearity and of neglecting an interaction are especially relevant for the case study analysed in this paper.

The base data generation process (DGP) considered for this Monte Carlo experiment consists of $N=1000$ binary Logit choices and 100 repetitions. The utility of alternatives depends on two attributes, x_1 and x_2, which are generated by taking draws from a random uniform distribution between zero and 2.5. The utility is completed by variable q^*_{in} and the error term e^*_{in}. As shown in Eq. (13) the utility coefficients for the three variables were -1 with no alternative specific constant.

$$U^*_{in} = -1x_{1in} - 1x_{2in} - 1q^*_{in} + e^*_{in} \tag{13}$$

The error e^*_{in} was defined as *iid* extreme value (0,1), and q^*_{in} was built as a convex combination ϕ of x_1 and an exogenous random uniform variable between zero and 2.5. By this, we were able to control the level of endogeneity in the DGP. When $\phi = 1$, q^*_{in} was equal to x_1 and when $\phi = 0$ q^*_{in} was fully exogenous. For the base case, ϕ was set to 0.2.

To correct for the endogeneity problem, two indicators are created. With reference to Eq (8)-(9), the first indicator is Iq_1, which is built as the sum of $\eta_1 q$ and random uniform variable between zero and one. The second indicator is Iq_2, which is built as the sum between $\eta_2 q$ and a random uniform variable between zero and one. Parameter η_1 is set to 1 and η_2 to 0.9 in the base case.

We begin analysing the impact of the sample size N on the performance of the MIS method in the correction for endogeneity. For this, we regenerate the base model considering, instead of 1000, subsequently 50, 100, 150, 200 and 250 observations. Results are summarized first at the upper left of Table 4 where we report the mean percent bias, across the 100 repetitions, of the estimator $\hat{\beta}_{x_1} / \hat{\beta}_{x_2}$ with respect to its population value 1. We compare the full model (including *q*), with the endogenous or curtailed model (omitting *q*) and the MIS model (omitting *q* and correcting with the MIS method). Results show strong evidence that sample size seems to have no impact on the performance of the MIS method. Indeed, the MIS performs worse in the context of small samples, but the same degree of small sample bias is observed for the full model. The same conclusion can be drawn from the upper left plot depicted in Figure 4 presenting box-plots of the sampling distributions of each method reported.

We also analysed the performance of the MIS method as a function of the degree of linearity of the relation between the indicator and the latent variable. According to Section 4, this potential problem may be present in the case study analysed in this paper. For this, we used a modified version of Eqs. (8) and (9), in which we allow the relation to be governed by a box-cox transformation where the value of *λ* determines the degree of linearity, as shown in Eq. (12). *λ* goes from 0 to 1. A *λ* close to 1 implies a linear model (as in Eqs. (8) and (9)) and as *λ* approaches zero the model takes the log form with respect to q..

$$Iq_{in} = \eta^{Iq}_{ss}\left[\frac{\left(q^*_{in}\right)^{\lambda} - 1}{\lambda}\right] + e^*_{ss\_in} \tag{13}$$

We explored the full range for *λ*, but only report the results for λ between 0.01 and 0.09 in Table 4 and Figure 4. In this range we observe the highest degree of non-linearity in the relation between the indicator and the latent variable. We also explored cases in which the indicator included in the utility, the instrument, and both indicators had a nonlinear relation with the endogenous variable in the application of the MIS method. For all cases, no noticeable impact in the performance of the MIS method could be attributed to the degree of linearity of the relation of the indicators with the endogenous variable.

The third experiment analyses the strength of the relation between the indicator and the latent variable. For this, we considered different values of η, which in this case were assumed to be the same for both indicators. The results reported in Table 4 and Figure 4 show that, indeed, when the indicator is weakly correlated with the latent variable, the MIS method could perform very poorly. Not only the variance of the estimator obtained with the MIS method is large (see Figure 4), but also the mean. However, the problem seems to be solved after some threshold. Although this problem is not present in our case study because the indicators are strongly correlated with the latent variable (see Figure 3), it should be remarked that this potential problem is relevant and may not be detectable in a RP case study.

Table 4: Monte Carlo Analysis of Mean % Bias of $\hat{\beta}_{x_1} / \hat{\beta}_{x_2}$ as Function of Model Shifts

% Bias by Sample Size				% Bias by Degree of Linearity			
N	True (benchmak)	Endogenous (curtailed)	MIS	λ	True (benchmark)	Endogenous (curtailed)	MIS
50	-32.5	51.5	-22.5	0.01	1.17	21.20	1.13
100	14.7	40.6	15.5	0.03	1.62	21.23	1.50
150	9.05	32.5	9.35	0.05	0.352	20.6	0.602
200	7.47	28.5	7.57	0.07	-0.653	19.0	-0.993
250	-0.411	19.9	-0.0461	0.09	0.234	20.32	0.0606
% Bias by Strength of Indicator				% Bias by Degree of Endogeneity			
η	True (benchmark)	Endogenous (curtailed)	MIS	ϕ	True	Endogenous (curtailed)	MIS
0.1	1.72	51.3	-124	0.1	1.53	11.5	1.5
0.2	1.92	52.2	-6.05	0.2	1.62	21.2	1.5
0.3	0.730	50.5	-0.898	0.3	-0.133	30.1	-0.0181
0.4	-0.151	49.2	-1.29	0.4	-1.08	38.6	-1.30
0.5	-0.659	49.5	-0.738	0.5	-0.659	49.5	-0.692

We then analyse the performance of the MIS method depending on the degree of endogeneity. For this, we modify the base model shifting the value of ϕ, which accounts for the degree of correlation between x_1 and the omitted q. In the experiment ϕ goes from 0.1 to 0.5 using a step size of 0.1. Results are summarized in the lower right of Table 4. As it occurred with the sample size, no noticeable impact in the performance of the MIS method seems to be attributable to the degree of endogeneity. Despite the mean percent bias grows with ϕ in the endogenous model, the performance of the MIS method each time is comparable to that of the full model. These results are displayed in the boxplots depicted at the lower right of Figure 4.

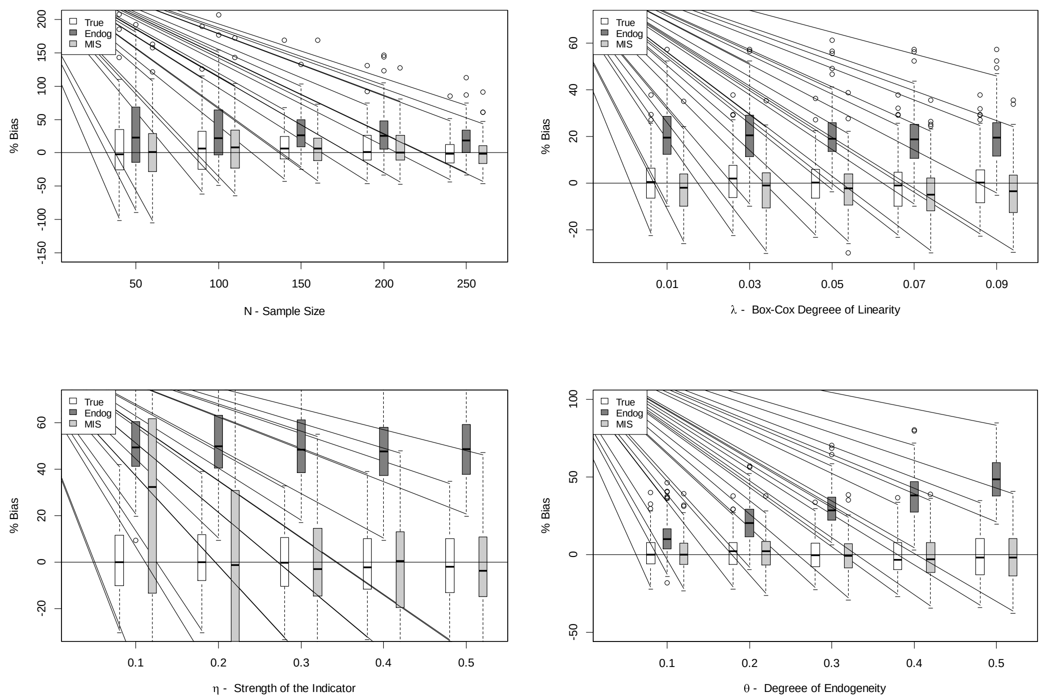

Figure 4: Boxplot of Monte Carlo Analysis of % Bias of $\hat{\beta}_{x_1} / \hat{\beta}_{x_2}$ as Function of Model Shifts

Finally, we examine the impact of neglected interaction. As discussed in Section 4, this problem seems to be present in the case study analysed in this paper and cannot be addressed with the indicators that are available. On the contrary, as Fernandez et al (2016) show, this issue may be solved if, by design, the indicators account directly for the interaction, as it may be the case for a RP experiment in which the interviewee is requested to assess a recent trip. A similar result would be obtained for a SP experiment in which the interviewee is asked to report their perception of an omitted attribute in each choice task.

We configure this experiment by adding a term $\gamma tt_{in} q_{in}$ to the utility of each alternative, varying γ from 0.01 to 0.09. As shown in Table 5 and Figure 5, the performance of the MIS method becomes quickly deteriorates as γ grows and becomes even worse than that of the uncorrected (endogenous) model.

Table 5: Monte Carlo Analysis of Mean % Bias of $\hat{\beta}_{x_1} / \hat{\beta}_{x_2}$ by Degree of Interaction

γ	True (benchmark)	Endogenous (curtailed)	MIS
0.01	0.317	19.8	0.424
0.03	1.99	17.1	-2.39
0.05	-0.0474	12.8	-5.54
0.07	-0.304	8.03	-10.0
0.09	3.03	6.20	-11.3

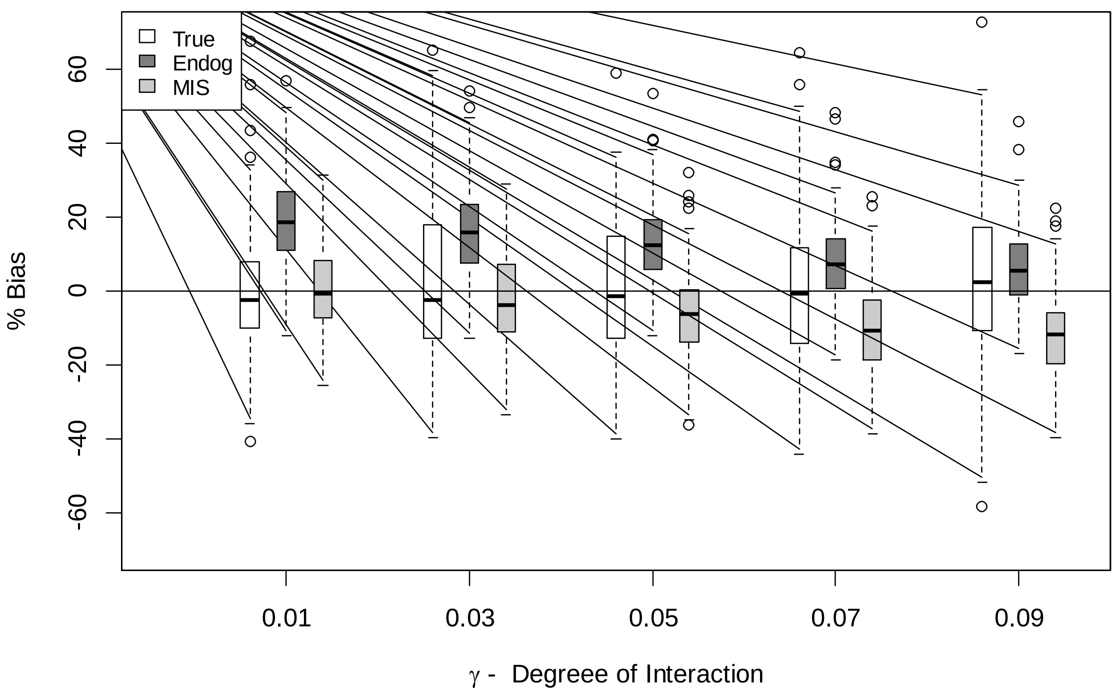

Figure 5: Boxplot of Monte Carlo Analysis of % Bias of $\hat{\beta}_{x_1} / \hat{\beta}_{x_2}$ By Degree of Interaction

From the Monte Carlo analysis, it can be concluded that the most likely source for the failure of the redundancy assumption detected in Section 4 is neglected interaction between the crowding level and observed model variables in the utility. More general, it is also shown that the strength of the instruments may also play an important role in the performance of the MIS method.

6. CONCLUSION

In this article we applied the Multiple Indicator Solution (MIS) method to correct for endogeneity due to omitted crowding in a discrete choice model of public transport. This is the first practical application of the MIS method to a public transport choice experiment illustrating its potential to become a valuable tool for practitioners by addressing a perennial estimation problem using simple ex-post questions to public transport users.

The case study consists of a Stated Preference (SP) survey where crowding conditions are illustrated by a picture, a diagram or a text description, and are then omitted from the choice model to artificially induce endogeneity. The MIS method applied here used Likert indicators of comfort and security provided by other individuals that were exposed to the same illustration of crowding levels.

Only the model that used Comfort as the instrument reached estimators statistically equal to those attained with a perfect (in the econometric sense) proxy, serving as the benchmark in our case study. It is found that this may be caused by the failure of the redundancy assumption, which in turn seems to be caused by a neglected interaction between the crowding level and travel time that was not properly picked up by the security indicator in the MIS comfort model.

The main contribution of the paper is the application of the MIS method to an SP experiment that was not specifically designed to account for it. This application mimics a real field experiment in which crowding conditions cannot be properly measured, but its perception can be collected by a post-trip questionnaire. More generally, the proposed approach has the potential to bypass the common problem of not being able of properly measure or estimate occupancy levels in public transport vehicles, and thus simplify the process to estimate the disutility of passenger crowding in public transport users. The practical relevance of the presented approach relies on the significant effect that passenger crowding has in public transport demand levels (Wardman and Whelan, 2011), route choice (Raveau *et al.*, 2011) and optimal public transport supply and pricing outputs (Tirachini *et al.*, 2014).

Throughout, we illustrate the way in which the MIS method can be applied in practice and present evidence about its potential and limitations. Using an extensive Monte Carlo analysis, neither the degree of nonlinearity in the relationship between the indicator and the omitted variable, nor the sample size, seems to play a significant role in the success of the MIS method. On the other hand, having indicators that are only weakly correlated with the omitted variable and neglecting an interaction effect can have a severe impact on the results. For the presented SP case study only the latter seems to have played some role in the results, but the problem is solved because one of the indicators picks up this ignored correlation with other policy attributes.

Regarding future lines of research on this area, it would be relevant to explore how to collect better indicators for omitted quality attributes in public transport, particularly in revealed preference experiments. The indicators used in this case study were originally collected with a different purpose and are not necessarily the best to address the issue at hand. For example, it could be argued that it may be better to ask passengers directly about their sensations and perceptions on how crowded the vehicle they just alighted from was; to record psychophysiological indicators of the passengers during their trips; and/or to infer the level of crowding from on-board pictures or carriage weighting tools. Another line of future research could be to assess the MIS method relative to the latent-variable approach using revealed preference data of this sort and using latent classes, random coefficients or systematic taste variations to account for the possible heterogeneity in the perception of the omitted attributes. Finally, a third possible line of research could be to investigate the degree of concordance and coherence of what respondents declare and what they experience, both in stated and revealed preferences experiments of public transport.

Acknowledgements

This publication was funded in part by CONICYT, FONDECYT 1150590 and by the Complex Engineering Systems Institute (CONICYT: FB0816). A preliminary version of this article was prepared when the first author was a Visiting Research Fellow at the Choice Modelling Centre of the Institute for Transport Studies of the University of Leeds, UK, with partial funding provided by a Leverhulme´s Visiting Professorship (VP1-Grant 2015-054).

REFERENCES

Aguiléra, V., Allio, S., Benezech, V., Combes, F., & Milion, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. **Transportation Research Part C: Emerging Technologies**, 43, 198-211.

Batley, R., Bates, J., Bliemer, M., Börjesson, M., Bourdon, J., Cabral, M. O., Chintakayala, P.K., Choudhury, C., Daly, A., Dekker, T., Drivyla, E., Fowkes, T., Hess, S., Heywood, C., Johnson, D., Laird, J., Mackie, P., Parkin, J., Sanders, S., Sheldon, R., Wardman, M., Worsley, T. (2017). New appraisal values of travel time saving and reliability in Great Britain. **Transportation**, forthcoming.

Cox, T., J. Houdmont and A. Griffiths, 2006. Rail passenger crowding, stress, health and safety in Britain. **Transportation Research Part A: Policy and Practice** 40(3): 244-258.

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. **Journal of Choice Modelling**, 20, 1-15.

Guevara, C. A. (2015). Critical assessment of five methods to correct for endogeneity in discrete-choice models. **Transportation Research Part A: Policy and Practice**, 82, 240-254.

Guevara, C. A. & Ben-Akiva, M. (2012) Change of scale and forecasting with the control-function method in logit models. Transport. Sci. 46 (3), 425–437.

Guevara, C. A. & Navarro, P. P. N. (2015). Detection of weak instruments when correcting for endogeneity in binary logit models. Proceedings of the 14th International Conference on Travel Behaviour Research (IATBR), Windsor, UK, July 19 - 23

Guevara, C.A. & Polanco, D. (2016) Correcting for Endogeneity Due to Omitted Attributes in Discrete-Choice Models: The Multiple Indicator Solution (MIS). **Transportmetrica A: Transport Science**, 12(5), 458-478.

Guevara, C.A., Ben-Akiva, M., 2006. Endogeneity in residential location choice models. Transport. Res. Rec.: J. **Transport. Res. Board** 1977 (1), 60–66.

Guevara, C. A., & Thomas, A. (2007). Multiple classification analysis in trip production models. **Transport Policy**, 14(6), 514-522.

Heckman, J., 1978. Dummy endogenous variables in a simultaneous equation system. **Econometrica** 46, 931–959.

Hausman, J., & McFadden, D. (1984). Specification tests for the multinomial logit model. Econometrica: Journal of the Econometric Society, 1219-1240.

Hensher, D. A., J. M. Rose and A. Collins (2011). Identifying commuter preferences for existing modes and a proposed Metro in Sydney, Australia with special reference to crowding. **Public Transport** 3(2): 109-147.

Jara-Díaz, S. R. and A. Gschwender, 2003. Towards a general microeconomic model for the operation of public transport. Transport Reviews 23(4): 453 - 469.

Karaca-Mandic, P. and K. Train (2003), Standard Error Correction in Two-stage Estimation with Nested Samples. **Econometrics Journal**, 6(2), 401-407.

Mohd Mahudin, N. D., T. Cox and A. Griffiths (2012). Measuring rail passenger crowding: Scale development and psychometric properties. **Transportation Research Part F** 15(1): 38-51.

OECD/ITF (2014). Valuing Convenience in Public Transport. ITF Round Tables, No. 156, OECD Publishing.

Petrin, A., Train, K., 2010. A control function approach to endogeneity in consumer choice models. J. Market. Res. 47 (1), 3–13.

Raveau, S., Muñoz, J.C., de Grange, L. (2011), A topological route choice model for metro. Transportation Research Part A: Policy and Practice 45 (2), 138–147.

Stock, J. H., and Yogo, M. (2002), "Testing for Weak Instruments in Linear IV Regression," NBER Technical Working Paper No. 284

Tirachini, A., D. A. Hensher and J. M. Rose, 2013. Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. **Transportation Research Part A** 53: 36-52.

Tirachini, A., D. A. Hensher and J. M. Rose, 2014. Multimodal pricing and optimal design of urban public transport: the interplay between traffic congestion and bus crowding. Transportation Research Part B 61: 33-54.

Tirachini, A., Sun, L., Erath, A., Chakirov, A., 2016. Valuation of sitting and standing in metro trains using revealed preferences. **Transport Policy** 47, 94-104
Tirachini, A., Hurtubia, R., Dekker, T., Daziano, R.A., 2017. Estimation of crowding discomfort in public transport: Results from Santiago de Chile. **Transportation Research Part A** 103, 311–326.
Train, K., & Wilson, W. W. (2008). Estimation on stated-preference experiments constructed from revealed-preference choices. **Transportation Research Part B: Methodological**, 42(3), 191-203.
Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. **Mathematical Social Sciences**, 43(3), 303-343.
Wardman, M. and G. A. Whelan, 2011. Twenty years of rail crowding valuation studies: evidence and lessons from British experience. **Transport Reviews** 31(3): 379-398.
Whelan, G. and J. Crockett (2009). An investigation of the willingness to pay to reduce rail overcrowding. International Conference on Choice Modelling, Harrogate, England, April.
Wooldridge, J. (2010), **Econometric Analysis of Cross-Section and Panel Data**, Second Edition, MIT Press, Cambridge, MA.