Transportation Research Part A 130 (2019) 752–769

Contents lists available at ScienceDirect

Transportation Research Part A

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

Comparing best-worst and ordered logit approaches for user satisfaction in transit services

Eneko Echaniz[a,*], Chinh Q. Ho[b], Andres Rodriguez[a], Luigi dell'Olio[a]

[a] University of Cantabria, E.T.S. Ingenieros de Caminos, Canales y Puertos, Avda. de los castros S/N 39005 Santander, Cantabria, Spain
[b] The University of Sydney Business School, Institute of Transport and Logistics Studies, 378 Abercrombie Street, Darlington, NSW 2008, Australia

ARTICLE INFO

Keywords:
User satisfaction
Importance
Public transport
Best-Worst scaling
Ordered logit

ABSTRACT

Customer overall satisfaction towards a public transport system depends mainly on two factors: how satisfied they are with different aspects that make up the service and how important each of the service aspects is to the customer. Traditionally, researchers use revealed preference surveys and ordered probit/logit models to estimate the contribution of each service attribute towards the overall satisfaction. This paper aims to verify the possibility of replacing the traditional method with the more cost-effective best-worst case 1 method, using a customer survey recently conducted in Santander, Spain. The results show that the satisfaction level obtained from these alternative methods are remarkably similar. The relative importance of each attribute delivered by the two methods differ, with the Best-Worst approach showing more intuitive and consistent results with the literature on public transport customer satisfaction. A regression method is developed to derive customer satisfaction with each service attribute from Best-Worst modelling results.

1. Introduction

Ordered logit and probit models have been widely used in public transport satisfaction studies (Alonso et al., 2018; Bordagaray et al., 2014; Echaniz et al., 2018). These models predict the overall quality of a transport service (i.e., overall satisfaction) based on the extent to which users are satisfied with each of the service attributes such as travel time, spatial coverage or service frequency. Thus, satisfaction data on each of the attributes that define the service have to be collected, usually based on a Likert scale on which respondents indicate their level of satisfaction with each service attribute and overall. Such surveys are usually lengthy and repetitive, resulting in low response rate or loss of sample due to respondent burden. For example, in a recent on-board surveys of bus users in Santander, Spain (Echaniz et al., 2018) many respondents were not able to finish the questionnaire based on traditional rating method. Consequently, interviewers had to leave the bus with the respondent in order to complete the survey. Not only reducing the interviewer's productivity but also complicating the logistics of survey.

An alternative approach to studying customer satisfaction and the contributions of different service attributes to the customer satisfaction is the Best-Worst (BW) case 1 method. Rather than asking the respondent to evaluate each service attribute at a time, as in the traditional rating method, the BW method shows the respondent a set of service attributes at the time and asks them to choose the best and the worst of the attributes shown and the process is repeated until all attributes are covered. Thus, BW surveys are less time consuming and more intuitive for respondents, requiring less guidance from the interviewer. Hence, BW survey method is a cost-

* Corresponding author.
E-mail addresses: echanize@unican.es (E. Echaniz), chinh.ho@sydney.edu.au (C.Q. Ho), andres.rodriguez@unican.es (A. Rodriguez), luigi.dellolio@unican.es (L. dell'Olio).

https://doi.org/10.1016/j.tra.2019.10.012
Received 9 January 2019; Received in revised form 22 October 2019; Accepted 22 October 2019
Available online 31 October 2019

efficient way to obtain data required for satisfaction studies, and yet this method has not been very used to study customer satisfaction in public transport systems, with only few exceptions (Beck et al., 2017; Beck and Rose, 2016).

This paper investigates the possibility of replacing the traditional approach to customer satisfaction with a best-worst case 1 survey method (Louviere et al., 2015). If this potential of BW method is verified, a more flexible and effective way to study customer satisfaction will be achieved, without imposing much burden on the respondent (through shorter questionnaires) or compromising the statistical robustness of the results. With respect to the latter, pioneer work of Beck and Rose (2016) suggests that the relative importance of service attributes identified by conventional rating methods can be biased. This limitation of the traditional method must be overcome, in Cao and Cao (2017) for example it was found that implicit importance (attribute importance derived from modelling estimates) based on ordered logit models is presumably a better indicator for decision making than the traditional rating. Motivated by these observations, this paper aims to compare the suitability of BW methods with that of traditional rating methods in studying customer satisfaction within the context of bus services.

The remainder of this paper is structured in to five sections. The next section reviews the relevant literature. Section 3 describes the data, the survey design and the sample. Section 4 explains the models, with estimation results presented in Section 5. The paper ends with the conclusions of the main findings, with some directions for future study.

2. Literature review

Budget constraints mean that investment options to improve services should be examined to maximise potential benefit for the given budget. One way to identify the most potential areas for improvement is to establish the extent to which users are satisfied with different service attributes. This can be achieved by asking users to give their level of satisfaction towards each service attribute through a traditional satisfaction survey with some ranking scale such as a 5-point Likert. However, investments should not be made based solely on a low level of satisfaction towards a certain attribute, since not all aspects of the service affect the users in the same way, with some service attributes being more important than others. A common solution in public transport (PT) customer satisfaction study is to ask the users how satisfied they are with the service alongside with how important each service attribute is to them. In doing so, the operator knows where to focus the improvement efforts to get the biggest benefit.

Customer satisfaction in PT services has traditionally been studied using revealed preference (RP) surveys. Different inception and interview methods have been used, such as on-board interviews (de Oña et al., 2013; Eboli and Mazzula, 2009; Eboli and Mazzulla, 2011; Echaniz et al., 2018), online surveys (Abenoza et al., 2017; Beck and Rose, 2016; Rose and Hensher, 2018) and mobile app (Guirao et al., 2015). Users evaluate different aspects of the PT service based on their experience using it, typically referred to as service attributes, based on qualitative or quantitative scales. These evaluations define the level of satisfaction users have with the different aspects of the PT service and also with the service bundle as a whole. The attributes to be evaluated by the respondent are usually derived from previous studies, as in the case of Hensher et al. (2003) where 13 representative items of the service were defined, or in Efthymiou et al. (2018) where factors that affect satisfaction are analysed for times of economic crisis. The subsequent analysis of the obtained data has been carried out in different ways over the years. The simplest and most popular method is descriptive analysis in which the mean and deviation of the level of satisfaction with each attribute and the entire service (i.e., overall satisfaction) are used to represent customer satisfaction. Thus, most studies on user satisfaction perform this basic analysis before applying a more complex and advanced method (e.g., de Oña et al., 2017; Eboli et al., 2018; Eboli and Mazzulla, 2015; Gonzalo-Orden et al., 2011; Tyrinopoulos and Antoniou, 2008).

The importance users place to different PT service attributes can be established using different methods, but they could be grouped into two approaches. The first approach uses the stated importance directly (explicit importance), where the customers rate each attribute on an importance scale similar to the one used for rating satisfaction. The second approach use derived importance in which the relative importance of each attribute is inferred by statistically analysing the relationship of individual attributes with the overall satisfaction (implicit importance) (Weinstein, 2000). Numerous studies were found to use the first approach to obtain the importance of the attributes (Beck and Rose, 2016; Guirao et al., 2016; Rose and Hensher, 2018) although this approach leads to an increase in the duration of the questioning, which means that surveys can become excessively long.

However, there are more efficient ways to obtain the importance level of the attributes, as is the case with the Best-Worst (BW) scaling (Louviere et al., 2015). Three types of BW exercises are used in the literature. These are the object case (Case 1), the profile case (Case 2), and the multi-profile case (Case 3). In Case 1, the respondent is asked to select the best and worst options from a series of objects or items such as a list of brand names, without showing any attribute or characteristic apart from the item itself. In Case 2, the respondent is asked to select the best and worst from a list of attributes, each of which was assigned a specific level determined by some experimental design. The choice is made between the different attributes with each one having their own set of levels. Case 3 is associated with the classic discrete choice experiments, where the choices are made between a set of alternatives composed of different attributes with different levels. The choice is then made between the different alternatives available which are composed by the same attributes but with different levels on each one. BW method have been successfully used in different transport studies. For instance, Beck et al. (2017) used a BW case 3 study to identify the respondent's attitudes towards choosing electric vehicles in the presence of regular fuelled alternatives. Or in Mulley et al., (2014) where BW case 1 was used to study the preference of the citizens regarding the construction of a Bus Rapid Transit (BRT) or a Light Rail Trains (LRT) service. In the context of public transport service

satisfaction, Beck and Rose (2016) applied also BW case 1 method and compared it with the traditional rating method. They concluded that the traditional way of establishing the importance of the variables is biased in which the service attributes people most satisfied with were associated with the highest levels of importance. In turn, it was observed that traditional responses did not show great variability. Beck and Rose (2016) also concluded that the traditional rating did not provide enough information for decision making. On the other hand, the BW method made it possible to better capture these variations. The correlation between satisfaction and importance was found to be much more coherent using the BW method, proving to be a more useful decision tool. Following Beck and Rose (2016) this study uses a BW case 1 method instead of a traditional method.

The analysis of the BW data has been made following two main techniques. In one hand, score measures i.e. analytical closed form (ABW) (Lipovetsky and Conklin, 2014) or normalized BW scores (NBW) (Louviere et al., 2015). On the other hand, discrete choice models have been applied, such is the case on Beck and Rose (2016) or Marley and Pihlens (2012). Most used discrete choice model is the Multinomial Logit. In Marley et al. (2016) an interesting comparison is made between ABW and NBW score measures, also, they compare the results obtained with a MNL model. The results show that for all three methods ABW, NBW and MNL are very close related.

Regarding the implicit importance, over the years different modelling methods have been used in order to find the most correct way to define the importance of the different attributes that define a service. de Oña and de Oña (2015) and delÍOlio et al. (2018) provide a comprehensive review of the methods. Recently, Allen et al. (2018b) and Allen et al. (2019) uses the structural equations model to study user satisfaction with Transantiago and Metro Madrid respectively. Mouwen (2015) performed an analysis of the public transport satisfaction in the Netherlands, the method used was the multiple regression. Discrete choice models have been also used, specially ordered models, both Logit and Probit (Allen et al., 2018a; Bordagaray et al., 2014; dell'Olio et al., 2010; Echaniz et al., 2018). Another way based on discrete choice modelling has been by carrying out stated preference (SP) surveys, which show the respondent a number of choice tasks and ask them to choose the one they most preferred. The data is analysed using discrete choice models of some kind. For instance, Román et al. (2014) used multinominal logit model (MNL) and Mixed Logit (ML) model to examine public transport services in Gran Canaria (Spain). Similarly, dell'Olio et al. (2011) used an MNL model to analyse the quality desired by future users.

Regardless of the data analysing method, the most statistically significant factors in most of these studies are the frequency of the service, the reliability and travel time, and to a lesser extent, the comfort of the buses and the smoothness of the ride. Allen et al. (2018a,b) showed that having a reliable service and a good frequency were the most influential attributes when explaining the users' satisfaction with the public bus system. The perceived waiting and travel times were also found to be of great importance. In addition, Mouwen (2015) showed that on-time performance, travel speed and frequency are most important attributes when explaining the overall quality of the service. Similarly, Román et al. (2014) showed that urban users has a greater willingness to pay for waiting time, travel time and access time. In dell'Olio et al. (2011) it was shown that in order to attract users it is necessary to increase the overall quality of the system by improving the waiting time, the comfort during the trip, the sources of information and the frequencies.

Thus, it appears that the majority of studies on public transport service satisfaction arrive at similar conclusions regarding the main drivers of customer satisfaction, even when using different modelling methods and datasets. Only one study was found that compares the results of an Ordered Logit model with a conventional rating based importance level (Cao and Cao, 2017). The main finding of that study was that the importance levels obtained with both methods were different. To the best of the authors' knowledge, no study has analysed the relationship between conventional rating and modelling with BW data focused in public transport satisfaction.

3. Data

3.1. Survey

The survey includes two parts. The first part asked respondents to report their socio-economic characteristics (e.g., gender, age, work status, income level), level of bus usage (e.g., trip purpose and number of trips per week) as well as the availability of alternative modes for these trips. The second part involved the respondent evaluating the importance and satisfaction towards different service attributes. A total of 24 service attributes, shown in Table 2, were used to define the services based on the existing international literature and several focus groups carried out in the city of Santander (Ibeas et al., 2011). These service attributes were grouped into six sets of four attributes each. Each respondent was asked to evaluate three sets of attributes allocated dynamically and randomly such that no attribute appeared twice for the same respondent and the total sample provides a balance of attributes assessed.

Each respondent was asked to evaluate the same set of attributes based on both traditional ranking (5-point Likert scale) and best-worst response mechanisms. In the traditional rating exercise, the respondent was asked to rate each attribute on a 5-point Likert scale (3 sets × 4 attributes/set = 12 attributes in total). In the best-worst scalling exercise, the respondent was asked to select, out of the same four attributes included in the choice task, which attribute they are most and least satisfied with (satisfaction choice), as well as which attributes are most and least important to them (importance choice). Fig. 1 summarise the data collected from one of the three choice tasks showed to each respondent. Respondents were not shown the three tasks of one question simultaneously but

Example	Rating scale (5 point likert scale)						Best-Worst choice			
	Very Bad	Bad	Normal	Good	Very Good	N/A	Most Important	Least Important	Most Satisfied	Least Satisfied
Attributte 1					X		X		X	
Attributte 2			X							
Attributte 3				X				X		
Attributte 4			X							X

Fig. 1. Example data collected from one choice task based on Rating vs. BW scales.

one after the other. At the end of the survey, all respondents were asked to rate the service as a whole, defined as Overall Satisfaction. The overall satisfaction was obtained following the same 5-point rating scale used for the attributes.

3.2. Sample

The surveys were run between October and November 2017 in the city of Santander. Face-to-face interviews were conducted on four bus lines operated by the municipal public company in the urban area of Santander. A total sample includes 808 completed interviews, spreading across the whole day with interviews taken place both at bus stops and on board.

Table 1 shows the main characteristics of the respondents. Women are over-represented in the sample (two in three respondents). A quarter of respondents are under 25 years old, while other age groups are more balance. Regarding occupation, almost half (47%) of the respondents were employed and nearly a quarter (24%) were students with the balance being retirees (17%), unemployed (8%)

Table 1
Respondents' socio-economic information.

Gender	Male	33%	
	Female	67%	
Age	< 25	25%	
	25–34	14%	
	35–44	15%	
	45–54	17%	
	55–64	15%	
	65–75	11%	
	> 75	4%	
Work status	Housekeeper	5%	
	Employee	47%	
	Unemployed	8%	
	Student	24%	
	Retired	17%	
Other available transport systems	Car (Driving)	35%	
	Car (accompanying)	12%	
	Bike	6%	
	Motorcycle	3%	
	Other	44%	
Trip purpose		Origin:	Destination:
	Home	46%	29%
	Work	22%	25%
	Studies	9%	13%
	Health	4%	5%
	Shopping	5%	7%
	Leisure	10%	13%
	Other	5%	9%
Number of trips made by bus per week	< 5	26%	
	5–15	54%	
	15–30	18%	
	> 30	1%	
Monthly income	< 900€	7%	
	900€–1500€	20%	
	1500€–2500€	17%	
	> 2500€	14%	
	No answer	42%	

Table 2
Satisfaction ratings.

Order	Attribute	Acronym	Mean	Mode	Std. deviation
1	Use of hybrid buses	HY	3.24	3	0.77
2	Access time to bus stop	AT	2.94	3	0.90
3	Egress time from alighting stop to destination	DT	2.91	3	0.89
4	Vehicle cleanliness	CL	2.81	3	0.70
5	Ease of transfer	TR	2.79	3	0.90
6	Information at stops	IS	2.76	3	0.97
7	Information on board	IB	2.73	3	0.90
8	Comfort of the buses	CM	2.71	3	0.74
9	Service reliability / punctuality	SR	2.70	3	0.86
10	Driver's kindness	DK	2.63	3	0.85
11	Quality of bus stops	ST	2.62	3	0.81
12	Information on mobile app	IM	2.61	3	1.27
13	Line coverage	LC	2.60	3	0.83
14	Information on the web page	IW	2.58	3	0.96
15	Priority seats for people with reduced mobility (PRM)	RM	2.51	3	0.89
16	Waiting time	WT	2.50	3	0.91
17	Crowding level	OC	2.50	3	0.87
18	Readability of map design	MD	2.48	3	0.98
19	In-vehicle travel time	TT	2.47	3	0.85
20	Service frequency and timetables	SE	2.44	3	0.97
21	Driving style	DS	2.39	3	0.86
22	Price/Fare	PR	2.33	3	0.94
23	Calefaction/air conditioning	CA	2.31	3	0.99
24	Noise	NO	2.28	2	0.82
	Overall satisfaction	OS	2.69	3	0.80

and housewives (5%). Half of the respondents had some other motorized alternative to make the same journey, while only a 6% would be willing to make the same journey by bicycle. The trips captured in the survey showed the important role of bus services in Santander for commuting purposes, i.e. travelling between home and office/school. Nearly half of the trips (46%) have the home as an origin and more than a quarter (29%) have the home as a destination. Work was the second main reason, both as an origin and a destination. The respondents are mainly habitual users with a low frequency of use per day. Specifically, more than half (54%) of the respondents use bus services up to 15 time per week. Finally, due to the sensitivity of the question, 42% of respondents decided not to answer the question related to their income level. Of the people who did answer, income levels have a good mix, with a greater proportion of people with an average salary: 20% of people between 900 and 1500 € per month and 17% between 1500 and 2500€.

3.3. Rating scale results

Table 2 shows the means of user satisfaction with each of the 24 service attributes in a descending order and the overall satisfaction. The traditional rating scale is based on a 5-point Likert scale and the rating is recoded to have the value from 0 (very unsatisfied) to 4 (very satisfied) for descriptive and modelling analysis.

Overall, the respondents are quite satisfied with the service, with an average rating of 2.69 out of 4. The service attribute that users are most satisfied with relates to the use of sustainable propulsion engines with hybrid vehicles (HY), being the only attribute that has the level of satisfaction exceeding 3. In total, nine attributes have a level of satisfaction greater than average, with the lowest level of satisfaction observed for noise level (NO), air conditioning (CA) and fare (PR).

4. Model specifications

This section describes the modelling approaches used to model customer satisfaction data obtained from the empirical survey. First, an Ordered Logit model is discussed and shown how it is used for modelling the data obtained from the traditional rating responses. This is followed by a specification of two standard logit models for the empirical data obtained from BW responses: one model for the level of satisfaction and another model for level of importance.

4.1. Ordered Logit for traditional raking data

Ordered Logit models are based on the following specification of a latent regression:

$$q_i^* = \beta' x_i + \varepsilon_i, \; i = 1, \ldots, n. \tag{1}$$

in which the latent continuous preference variable q_i^* is only observed in discrete form q_i through a censoring mechanism:

$$\begin{aligned} &q_i = 0 \; if \; q_i^* \leqslant \mu_0 \\ &q_i = 1 \; if \; \mu_0 < q_i^* \leqslant \mu_1 \\ &\cdots \\ &q_i = J \; if \; \mu_{J-1} < q_i^* \leqslant \mu_J \end{aligned} \tag{2}$$

Note that the specification in Eqs. (1) and (2) assumes that neither parameters $\boldsymbol{\beta}$ nor thresholds μ vary across individuals. This assumption of homoscedasticity is arguably strong and can be relaxed. The vector $\boldsymbol{x_i}$is a set of K covariates that are assumed to be independent of ε_i;and $\boldsymbol{\beta}$is a vector of K parameters to be estimated, together with $J + 2$ threshold parameters μ_j using N observations. The assumption of the disturbance ε_i completes the model specification. The conventional assumptions are that ε_i is continuous, random and follows a certain cumulative distribution function (CDF), $F(\varepsilon_i|\mathbf{x}_i) = F(\varepsilon_i)$.

For this study, q_i^* represents the non-observable overall satisfaction of the PT service, while q_i is the observable overall satisfaction obtained from the traditional rating question asked at the end of the survey. J represents the 5-point Likert scale options of the rating task shown in Fig. 1; $\mathbf{x}_i$ are the satisfaction ratings of the service attributes assessed by the respondent i with values ranging between 0 (“Very Bad”) and 4 (“Very Good”).

The dependent variable of the model is defined as the overall satisfaction (OS), while the independent variables are the satisfaction levels of the different attributes of the service. In total 24 independent variables or service attributes have been defined; however, conducting on-board face to face surveys means that a large proportion of bus users would not have enough time to provide their level of satisfaction towards each of the 24 attributes. Thus, each respondent was asked to rate 12 of the 24 attributes, which generates an additional modelling challenge but that can be solved using imputation methods to complete the database. The method used to complete the sample has been based on Multiple Imputation (Rubin, 1978, 1977), explained below. Echaniz et al. (2019) have shown that it is possible to obtain Ordered satisfaction models by using this method and that results obtained using this method are similar to those obtained with a complete database.

Multiple imputation is estimated by using a procedure called Fully Conditional Specification (FCS), which uses an iterative Monte Carlo method with Markov chains (van Buuren, 2007). The FCS approach is based on variable-by-variable imputation of data, specifying an estimation model for each one of the variables with missing data. The FCS tries to define $P(X, C, R|\theta)$ by specifying a conditional density $P(\mathrm{X_i}|C, \mathrm{X_{-i}}, R, \theta_i)$ for each $\mathrm{X_i}$, this density is used to impute X_i^{mis} given some C, X_{-i} and R. An imputation consists of a complete cycle through all X_i (van Buuren, 2007). Where X represents the evaluation of the attributes, C the characterization variables, θ the parameters of the imputation model and R an indicator that show if X is a missing or observed value. The imputation is made by using the Gibbs sampling methodology (Casella et al., 2016; Gilks et al., 1996) assuming that the conditional density distribution exists. This methodology has been used in a large number of simulation studies (Brand, 1999; Horton et al., 2016; Raghunathan et al., 2001; Van Buuren et al., 2006) that have provided sufficient evidence that the results obtained through the FCS are generally unbiased and have adequate coverage.

Once the database is completed, the models are estimated as usual. To estimate the model some normalizations are required. First, to keep the positive sign of the probabilities it is required to $\mu_{j+1} > \mu_j$. As the variable q_i^* exists in the entire real line and the model contain a constant term $\beta_0 = 0$, it is necessary to define $\mu_0 = 0$ and $\mu_J = +\infty$. As. The data does not contain information about the scaling of the dependent variable q_i^*, therefore, the free variance parameter $Var(\varepsilon_i) = \sigma_\varepsilon^2$ cannot be estimated. The usual approach is to assume that σ_ε is constant and depends on the distribution assumed for ε_i. In Logit models it is assumed that ε_i follows a logistic distribution, resulting in $Var(\varepsilon_i) = \pi^2/3$. The associated probabilities are defined as:

$$Prob[q_i{=}j|x_i] = Prob[\varepsilon_i \leqslant \mu_j - \beta' x_i] - Prob[\mu_{j-1} - \beta' x_i] > 0, j = 0, 1, \ldots, J; \tag{3}$$

The model (3) is estimated using maximum likelihood estimator which maximises the log-likelihood function defined as follows:

$$\log L = \sum_{i=0}^{N} \sum_{j=0}^{J} m_{ij} \log[F(\mu_j - \beta'_k x_{ki}) - F(\mu_{j-1} - \beta'_k x_{ki})] \tag{4}$$

where $F(\cdot)$ is the cumulative distribution function; $m_{ij} = 1$ if $q_i = j$ and 0 otherwise.

4.2. Multinomial Logit for Best-Worst scaling

The literature defines three type of BW data, known as Case 1, Case 2 and Case 3 as reviewed in Section 2. Aiming to verify the importance of different service attributes associated with bus services without including different attribute levels, this paper adopts the BW Case 1 method.

There are a total K attributes to be chosen on the survey. In each BW task a subset Y of four attributes is shown. With the answers of the choice task, a vector $\delta = (\delta_1, \cdots, \delta_k)$ is estimated, which is the utility coefficient of each attribute.

The probability of choosing an attribute $b \in Y$ as best is denoted as $P_B(b|Y)$. In the same way, the probability of choosing an

attribute $w \in Y$ as worst is denoted as $P_W(w|Y)$. The *joint* probability of choosing attribute b as best and attribute $w \neq b$ as worst is defined as $P_{BW}(bw|Y)$. In the experiment, the respondent had to select the best option and the worst option out of a subset of service attributes. The survey instrument was programmed in such a way that the respondent cannot advance if the same attribute was selected as both best and worst attributes. That is, the probability of the same attribute - call it x - being chosen as best *and* worst by the same respondent i (suppressed from the notations for simplicity), is always 0. Mathematically, $P_{BW}(xx|Y)=P_B(x|Y)P_W(x|Y) = 0$ since either $P_B(x|Y)$ or $P_W(x|Y)$ or both must be 0.

Adopting a standard logit specification to describe the choice of the best and the worst attributes, (i.e. assuming that the unobserved components of the utility follow Type 1 Generalized Extreme Value or Gumbel distribution with random variables independently and identically distributed), the probability $P_{BW}(bw|Y)$ for one BW choice task is defined as (5), which also is called the Maxdiff model (Marley and Louviere, 2005):

$$P_{BW}(bw|Y) = \frac{\exp[v(b) - v(w)]}{\sum_{\substack{l,k \in Y \\ l \neq k}} \exp[v(l) - v(k)]} \tag{5}$$

where $v(\cdot)$ is the observable utility components specified as a linear-in-parameter function of attributes such as $v(k) = \delta_k y_k$ where y_k is an indicator vector of 0 and 1 ($y_k = 1$ when the attribute k is shown to the respondent i and 0 otherwise. In this way, the parameter estimate δ_k could be interpreted as the importance or satisfaction level (depending on which model is being analysed) of attribute k relative to the reference/base attribute which has $\delta_0 = 0$.

The maxdiff model assumes that the respondent simultaneously chooses the best and the worst options; however, it may be possible that the respondent selects the best option first, then eliminate this attribute out of the choice set before selecting the worst option. In this case, the repeated best-worst model specification as in Eq. (6) may be more appropriate (Dyachenko et al., 2012).

$$P_{BW}(bw|Y) = P_B(b|Y)P_W(w|Y - \{b\}) = \frac{\exp v(b)}{\sum_{l \in Y} \exp v(l)} \cdot \frac{\exp - v(b)}{\sum_{k \in Y - \{b\}} \exp v(k)} \tag{6}$$

Similarly, if we assume that the respondent selects the worst option first, then the best option, a repeated worst-best model as in Eq. (7) could be used (Dyachenko et al., 2012).

$$P_{WB}(wb|Y) = P_w(w|Y)P_b(b|Y - \{w\}) = \frac{\exp v(w)}{\sum_{l \in Y} \exp v(l)} \cdot \frac{\exp - v(w)}{\sum_{k \in Y - \{w\}} \exp v(k)} \tag{7}$$

Table 3
Estimation results of BW and ordered logit models (t-value in parentheses).

Variable	MNL- BW (Satisfaction)	MNL - BW (Importance)	Ordered Logit
Use of hybrid buses	1.326 (10.96)	1.083 (8.66)	0.039 (1.87)
Access time to bus stop	1.105 (9.34)	1.46 (11.88)	0.151 (2.80)
Egress time	1.054 (8.9)	1.630 (13.1)	0.067 (3.07)
Vehicle cleanliness	0.794 (6.84)	0.890 (7.36)	0.108 (2.29)
Ease of transfer	0.914 (7.84)	1.161 (9.41)	0.134 (4.24)
Information at stops	0.860 (7.27)	1.319 (10.65)	0.155 (2.86)
Information on board	0.455 (3.91)	0.249 (2.01)	0.059 (2.68)
Comfort of the buses	0.666 (5.71)	1.112 (9.02)	0.063 (2.63)
Service reliability/punctuality	0.763 (6.62)	2.336 (18.6)	0.270 (4.56)
Driver's kindness	0.480 (4.18)	0.624 (5.07)	0.021 (1.98)
Quality of bus stops	0.284 (2.41)	0.584 (4.67)	0.154 (2.76)
Information on mobile app	0.585 (4.93)	0.958 (7.68)	0.045 (2.23)
Line coverage	0.548 (4.67)	2.235 (17.56)	0.200 (3.59)
Information on the web page	0.494 (4.4)	0 (0)	0.101 (3.35)
Priority seats for PRM	0.406 (3.45)	1.621 (12.89)	0.055 (2.47)
Waiting time	0.375 (3.21)	1.988 (15.76)	0.142 (3.14)
Crowding level	0.209 (1.81)	1.478 (12.03)	0.145 (3.28)
Readability of map design	0.114 (0.98)	1.118 (8.97)	0.180 (3.25)
In-vehicle travel time	0.434 (3.66)	2.239 (17.29)	0.080 (3.71)
Service frequency and timetables	0.305 (2.62)	2.260 (17.82)	0.298 (5.11)
Driving style	0.075 (0.65)	1.474 (11.8)	0.078 (2.41)
Price/Fare	0 (0)	1.860 (15.03)	0 (0)
Calefaction/air conditioning	0.050 (0.43)	0.430 (3.49)	0.062 (2.86)
Noise	0.069 (0.6)	0.472 (3.87)	0.030 (2.66)
Constant	–	–	−1.305 (−2.88)
Mu(01)	–	–	1.427 (11.07)
Mu(02)	–	–	3.628 (33.61)
Mu(03)	–	–	6.997 (46.47)
Log-likelihood	−5817.674	−5373.256	−818.697
AIC/N	4.819	4.452	2.072
McFadden PseudoR2	–	–	0.11

Eqs. (6), (7) are alternative model specifications of equation (5). Empirical study (see for example Greene, 2016; Ho and Hensher, 2017) shows that these alternative model specifications are likely to produce similar results. Thus, this paper uses the Maxdiff model and acknowledges that the repeated best-worst or repeated worst-best specifications could be used as an alternative specification.

5. Results

5.1. Discrete choice models

Three models were estimated using Nlogit v6.0: two MNL models based on the BW data obtained from the BW choice exercises and one OL Table 3 shows the parameter estimates with t-values shown in parentheses. The BW models show the order of the attributes in terms of satisfaction/importance levels. In each model, the parameter associated with the attribute that has the lowest level of satisfaction/importance is set at 0, allowing all other parameters to be positive, assisting parameter interpretation. Specifically, the value of a parameter identifies its position in the satisfaction/importance scale with the larger parameter being representative of a more satisfied/important attribute.

The remaining column show the values of the parameters of the OL model. The OL model is completed with the constant term and the threshold parameters, as was explained in Section 4. The dependent variable for this model is the overall satisfaction (OS) and the independent ones are all the attributes evaluated using the conventional rating scale. The values of the parameters in the OL model show how much each attribute contribute to explain the customer overall satisfaction with the service (i.e. OS).

For brevity, estimation results from the standard logit and order logit models are analysed in this article. These models assumes homogeneity in preference among individuals by estimating average parameters for the sample. This homogenous assumption is relaxed by estimating random parameter (RP) models (estimation results shown in the Appendix A). For brevity purpose, the comparison of alternative modelling approaches is conducted using standard models since the main findings, discussed below, still stand in light of the RP modelling results.

As stated in the BW satisfaction model, users are very satisfied with the company's environmental policy when deploying hybrid buses (HY). In addition, users are also very satisfied with the access times to the stops (AT) and the egress times (DT), this mean that bus users in Santander see a good spatial coverage. As for the less satisfactory attributes, the price of tickets (PR) can be found as the least satisfactory. The current tickets fares vary depending on user type and the payment system used; however, the price is not higher than other public transport services nearby. The result may suggest the existence of a strategic voting behaviour, in the sense that the respondents strategically voted down their satisfaction with transport fares to reduce the chance that operators may increase fares in the future. This behaviour was also observed in previous studies conducted in the same city (Echaniz et al., 2018). The environmental characteristics such as noise (NO) and heating systems (CA) also show low levels of satisfaction, as well as several attributes related to comfort during the trip: driving style (DS) and crowdedness (OC). Analysing the existing information channels in the service, it can be observed that the users are very satisfied with the information offered at stops, and somewhat less with the information available in mobile applications. While it is shown that for the remaining information sources (information on board the buses and on the website) users are not satisfied.

The importance BW based model shows that the most important attributes are those directly related to the basic characteristics of the service such as service reliability (SR), frequency (SE), on-board travel time (TT) and coverage of the lines (LC). Conversely, the least important attributes are noise level (NO), air conditioning systems (CA), information on board (IB) and on the website (IW).

According to the OL model, the attributes that show the high parameter values are service frequency (SE), service reliability (SR) and coverage of the lines (LC). By contrast, the driver kindness (DK), noise level (NO) and the use of hybrid technologies (HY) are the ones with the lowest parameter estimates. The threshold parameters show a nonlinearity in different rating points, which means that from the user's perspective, difference levels of effort are required to improve the service by one satisfaction point, such as from very bad to bad vs. from good to very good. These results are consistent with the findings obtained in the previous study developed in the same city (Echaniz et al., 2018).

With the data available we wanted to analyse if there is any connection between the models derived from the Best-Worst exercise and the results obtained from the Ordered Logit modelling and satisfaction ratings. Table 4 shows the correlation in parameters obtained from the two approaches. The correlation between the averages of the traditional satisfaction rating for each attribute and the BW satisfaction model is nearly perfect, with correlation coefficient of 0.95. The ordered Logit model shows a considerable correlation ($r = 0.486$) with the BW model of importance.

A deeper investigation into these strong correlations is presented in Figs. 2 and 3. Parameter values differ in scale from one model to another, and thus a direct comparison of parameter estimates does not show the true correlation between them. Therefore, both

Table 4
Correlation coefficients.

	Importance_BW	Satisfaction_BW	Ordered Logit	Satisfaction Rating
Importance_BW	1			
Satisfaction_BW	0.083	1		
Ordered Logit	0.486	0.074	1	
Satisfaction Rating	−0.056	0.946	0.016	1

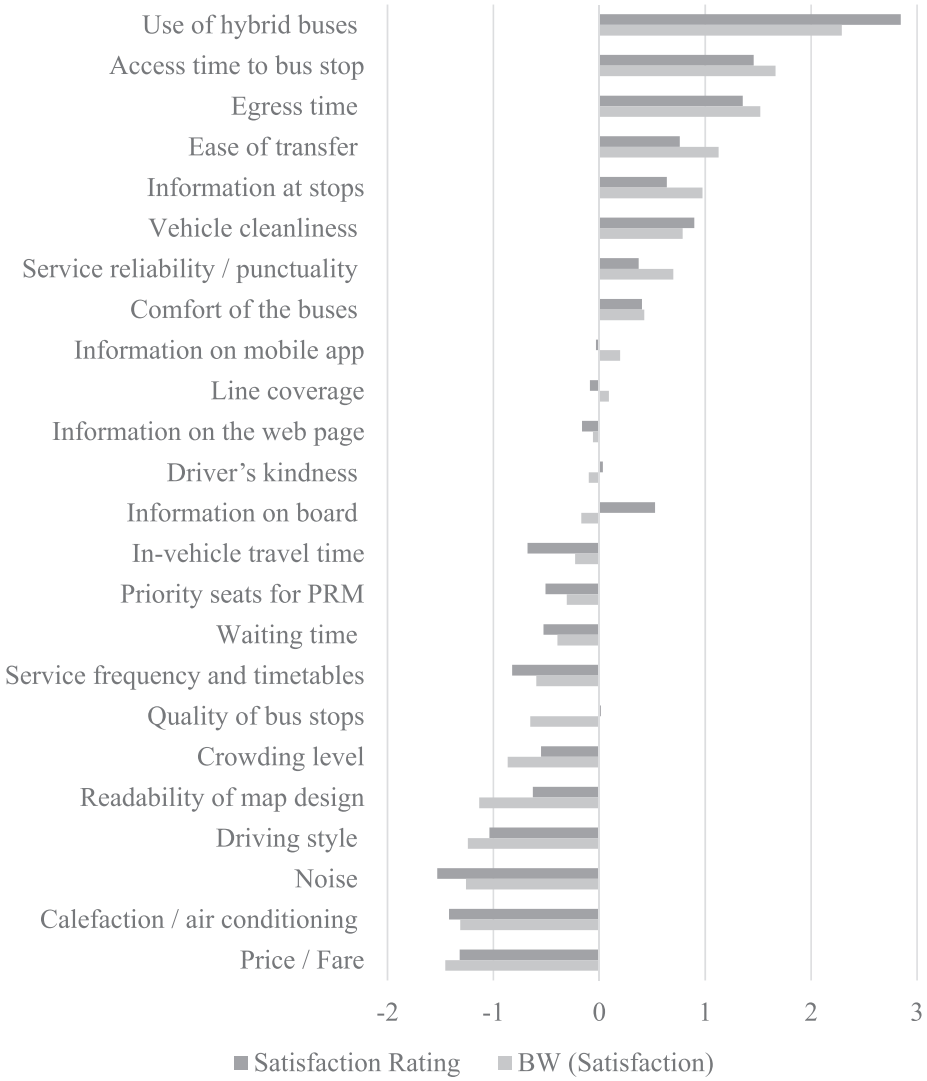

Fig. 2. Comparison between BW (satisfaction) model and satisfaction rating (OL).

sets of the parameter estimates were standardized, providing some positive and some negative standardized scores showed in Figs. 2 and 3. The first comparison has been made between the two values representing the satisfaction. On the one hand, we have the information based on the classic revealed preference survey rating, obtaining the average satisfaction with values between 0 and 4. On the other hand, the MNL model based on the BW satisfaction data. As can be seen in Fig. 2, the correlation between these two values is considerably high. Most of the attributes show a similar tendency for both values although there are a few exceptions such as information in mobile apps (IM), line coverage (LC), information on board (IB) and quality of the stops (ST). Therefore, both methods lead to the same results, lending support to the hypothesis that BW method can replace the traditional satisfaction rating.

For the comparison of the importance, the Ordered logit and MNL models based on the importance BW have been selected. The value that a parameter has in an Ordered model can be associated with the weight it has when explaining the dependent variable. In other words, the parameter explains the contribution of each attribute to the overall satisfaction. An analysis of these two set of parameters, presented in Fig. 3, shows a lower level of correlation than previous satisfaction one. Even the level of importance of the most important attributes shows a similar trend in both models, the rest of the attributes show very small correlation. In consequence, although both models represent a certain level of importance of the variables, these values are not the same and, therefore, represent different importance concepts.

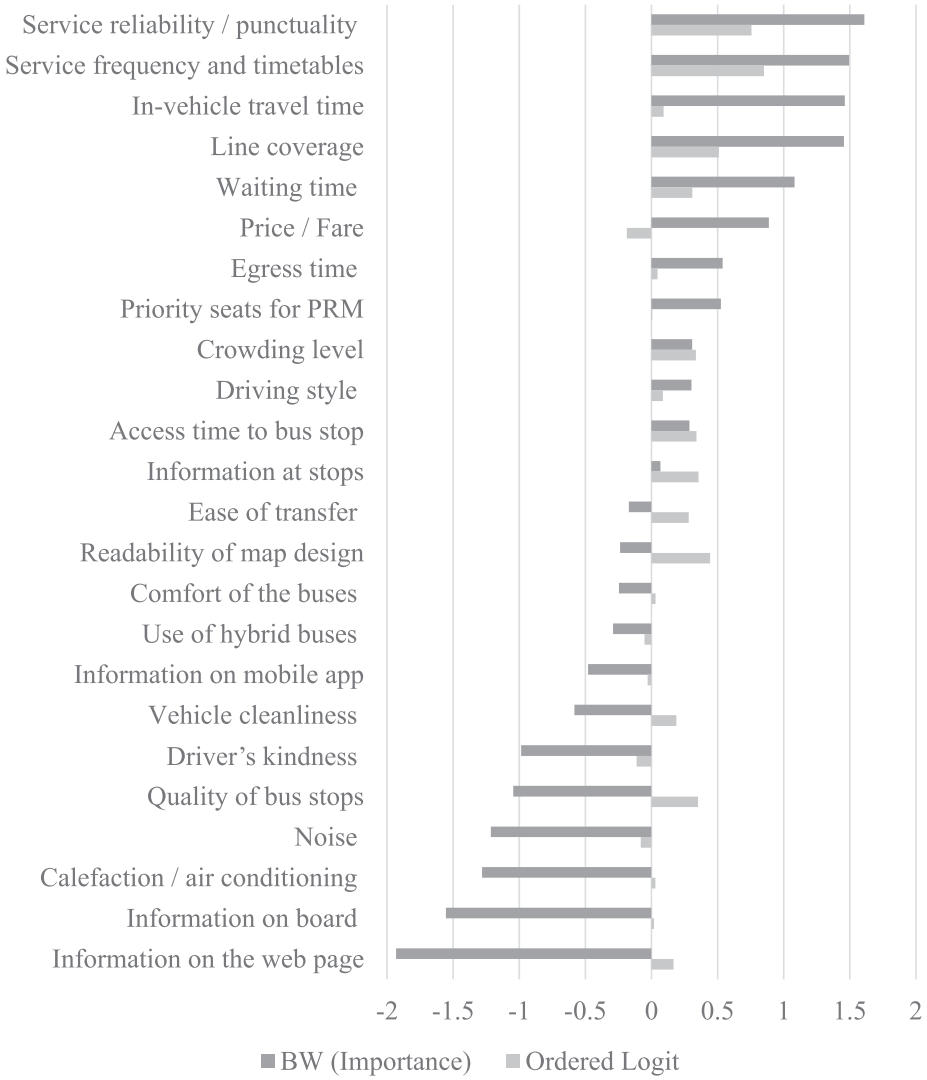

Fig. 3. Comparison between BW (importance) and Ordered Logit models.

5.2. Importance-performance analysis

The importance-performance analysis (IPA) (Martilla and James, 1977) is a widely used decision tool. The basis of this method is to cross both result (performance level and importance) in the same graph. Four quadrants are defined, each of them with a different level of improvement priority. The four quadrants are typically identified as 'keep up the good work' (Q1 – "Important and satisfied"), 'possible overkill' (Q2 – "Not important and satisfied"), 'low priority' (Q3 – "Not important and not satisfied") and 'concentrate here' (Q4 – "Important and not satisfied") (Sever, 2015). The attributes on the Q1 are considered the strengths of the service, attributes that are performing well and where investments should be kept equal to maintain the satisfaction level. Attributes on the Q2 contain attributes that are not important for users but still are performing strongly, that means that there is a possible waste of resources used in these attributes. Attributes on the Q3 are the ones with the lowest level of priority for investment. Finally, Q4 shown the main improvement priorities of the service, attributes that are important for users but are not performing good enough to satisfy the customers.

IPA method was recently used in the literature to compare explicit and implicit importance on transit services (Cao and Cao, 2017). Explicit importance is obtained by asking the respondent, either through a conventional rating method or some other methods, while implicit importance is derived from an OL model. They found that the priorities for service improvement based on explicit importance are different from those based on implicit importance. Aiming to verify this important finding, Fig. 4 presents the

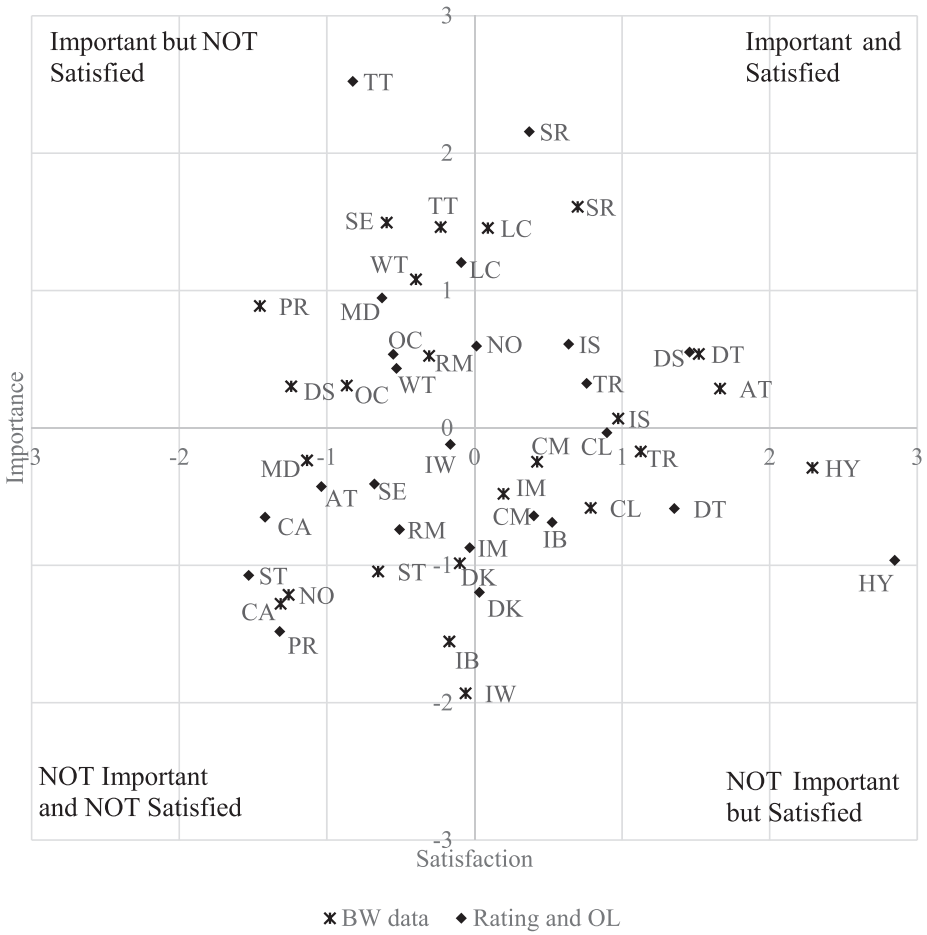

Fig. 4. Importance – performance analysis.

Importance – Performance Analysis (IPA) in which explicit importance and satisfaction obtained from BW model are compared with those obtained from the conventional method (rating and OL). Satisfaction levels are represented in the horizontal axis, while importance levels are shown on the vertical axis. Blue dots show the positions of the service attributes according to the BW data and modelling results. Orange dots show the same but using the conventional rating as satisfaction and OL parameters as importance. The values are normalized for presentation purposes.

Close to 55% of the service attributes position themselves in a different quadrant when using a conventional satisfaction rating vs. best-worst rating. For some attributes, such as line coverage (LC) and information on mobile phones (IM), the quadrants are different but positions are very close to each other. For some other attributes, such as ticket fare (PR), the differences are quite remarkable in which the BW method identifies ticket fare as an important variable while the conventional OL model suggests the opposite (PR is the not important).

The differences are much greater in importance than in satisfaction. Moreover, satisfaction values are quite similar between the two methods, according to the positions relative to the horizontal axis. Most of the attributes are placed in a very close horizontal axis value, which means a similar satisfaction level. Quadrant position differences are therefore a result of different important levels identified by the alternative methods. Given that attribute importance derived from BW model is explicit (since users explicitly choose the most and least important attributes from a set of attribute) and that of the OL model is implicit (deriving from parameter estimates of an overall satisfaction model), we support Beck and Rose (2016) argument that BW scaling is better than conventional rating for defining attribute importance.

Using results of the BW methods, a change to the current fare structure may result in a higher level of satisfaction since PR is highly important but is the most unsatisfied attribute. However, Results of the ordered model, suggests that any change in the fare policy would not generate increase the overall satisfaction of the service. Such the conflicting evidence is applied also to the service frequency in which the BW method suggests improving service frequency would improve customer satisfaction while the OL method identifies service frequency as a low priority. These differences suggest that the concept of importance may differs between the two

Table 5
Satisfaction regression model.

	Parameters	t-value
Overall Satisfaction	0.774	52.017
exp[BW(Satisfaction)]	0.301	14.397
R^2	0.999	
Adjusted R^2	0.954	

Table 6
Importance regression model.

	Parameters	t-value
Satisfaction ratings	0.017	1.914
Exp[BW(Importance)]	0.015	3.282
R^2	0.788	
Adjusted R^2	0.733	

modelling methods. We further investigate this in the next section.

5.3. Deriving attribute-specific satisfaction from BW models

An outstanding question regarding the use of BW survey method in customer satisfaction study is that whether it would be possible to obtain customer satisfaction level with each service attribute from the BW survey methods. This section addresses this question by performing a regression analysis to estimate the average level of satisfaction with each attribute based on the overall level of satisfaction and the BW model parameters. Two models are estimated: one for satisfaction level and one for the importance level of each attribute.

The BW satisfaction model shown in Section 5.1 represents the relative satisfaction level of the many attributes that together define the entire service. Thus, these parameters are best interpreted as how much more or less an average customer is satisfied with a certain service attribute, compared to the reference attribute. That challenge is to convert this relativity of satisfaction level to an average level of satisfaction for each attribute, which is usually available in the traditional rating survey with specific questions. We propose a way to supplement this information for BW method by adding a constant term to the regression model, effectively converting the BW model parameters to the satisfaction level for each of the attributes included in the BW survey. More specifically, it is the Overall Satisfaction that acts as the constant term, as it remains constant for the whole sample. The regression model is specified in equation (8) where the dependent variable is the satisfaction rating for each attribute and the independent variables are the Overall Satisfaction (OS) and the BW satisfaction model parameters.

$$SatisfactionRating_i = OverallSatisfaction + e^{\delta_k^{\mathrm{BW(Satisfaction)}}} \tag{8}$$

Table 5 shows the estimation results, confirming that it is possible to accurately estimate the average satisfaction level for each of the attributes from the BW data. The model has an R^2 of 0.999, indicating that nearly 100% of the variation in attribute-specific satisfaction is explained by the OS level and BW model parameters (i.e., the relativity satisfaction of different attributes). The results suggest that conventional rating surveys can be replaced by BW scaling surveys for measuring attribute satisfaction.

The parameters associated to an Ordered model define up to a certain level the implicit importance of the different attributes. As shown in Section 5.1, the implicit importance of the OL model and the explicit importance derived from the BW MNL model are correlated, although the correlation is not as strong as in the case of satisfaction. Table 6 shows the result of the regression model defined in Eq. (9).

$$\beta^{OL} = SatisfactionRatings_i + e^{\delta_k^{\mathrm{BW(Importance)}}} \tag{9}$$

The goodness of fit indicators shows that the parameters of the OL model can be estimated using BW data to a certain extent. This means that there is a difference in how importance is captured in each method. In BW method, importance is considered explicit while in OL it is implicit, and thus, they are related but not the same. The parameters of an OL model encapsulate both the explicit importance and the satisfaction. In other words, the parameters of the OL model represents not only the importance that each attribute as a separate factor has within the system, but also the satisfaction with the entire system. This explains why there is such a difference in the IPA analysis shown in Section 5.2, as the OL goes further than simply defining a level of importance of the different attributes.

6. Conclusions

This study showed that the conventional method (rating and ordered logit models) and the BW methods are both suitable for analysing users' satisfaction with public transport services. Correlation and modelling analysis results indicates that conventional rating and BW methods are equivalent when studying the satisfaction levels of the different attributes of the service. The regression model shows that it is possible to reproduce the rating results by using BW data, and thus, BW method can replace the conventional rating method. This is an important finding which has a significant implication for improving the efficiency of customer satisfaction surveys and bringing positive effect to the respondents. With the BW survey methods being less time consuming and easier for the respondent to answer than the traditional rating method, this finding suggests that we can replace the lengthy and repetitive customer survey with a series of games presented as best-worst choice tasks, especially when the results reported in this paper could be replicated on different datasets and/or in different settings.

The roles that different service attributes play in explaining customer overall satisfaction turn out to be very different, depending on the modelling methods. This finding is consistent with the growing evidence on the difference between explicit importance and implicit importance in customer satisfaction studies. Specifically, important drivers of customer satisfactions obtained from the best-worst explicit importance levels are travel time, service frequency and price while these key service attributes did not come out as important factors based on the OL implicit importance levels. In this sense, the BW scaling appears to be a good indicator of attribute importance, while the OL model parameters goes further than just defining an importance level. The regression model has shown that OL model parameters are not only influenced by the importance of the attributes but also by their satisfaction level.

The Importance-Performance Analysis (IPA) offers a new way to classify service attributes according to their importance/satisfaction levels. This helps transport operators and authorities to identify the key service attributes to improve. Again, the conclusions can be considerably different and sometimes opposite, depending on the adopted method. The main differences between the two methods are ticket fare (PR) and service frequency (SE). The main differences are placed in importance levels, as satisfaction results are very similar between both methods. The importance levels derived from BW data are in line with the literature. In conclusion, the explicit importance levels obtained by using the BW method are more accurate than the implicit importance derived from the OL model. Therefore, the IPA based on BW scaling is a better indicator of which attributes should be the real priorities for operators. This finding verify results of Cao and Cao (2017), who concluded that improvement priorities based on implicit importance (OL model) were more reliable than those based on the explicit conventional rating.

Although the results of this study show that BW scaling can effectively replace the conventional rating method, several considerations should be taken into account when implementing the BW scaling method for customer satisfaction surveys. First, to obtain the average satisfaction rating of the attributes by using BW data, it is necessary to fit the regression model that connects both models, therefore, a preliminary study is required to estimate the regression model, similar to the study carried out in this article. In addition, the Overall Satisfaction of the service must be rated independently to the method used.

Regarding to future research, results reported in this paper suggest that correlations between the means of random parameter estimates are as strong as correlations between non-random model parameters; however, there are differences in the deviations of the random parameters in the sense that less preference heterogeneities were found in the BW data than in the rating data. This might suggest that BW survey method results in less noise in the data than the traditional rating survey method does. More research is required to verify this initial finding, such as using different datasets and/or parsimonious models where preference heterogeneities are segmented into systematic vs random heterogeneity. Also, different modelling approaches should be tested to fit the BW data, for example the repeated best-worst model. In addition, in an ongoing research we are investigating the extent to which customer satisfaction varies across the service levels during the whole day. Automatic Vehicle Location data is being used to identify not only the line but also the bus the respondents were on.

Acknowledgments

We greatly appreciated the three anonymous referees' constructive comments which have materially improved the paper. This study has been possible thanks to the financing of the Spanish Ministry of Economy and Industry in the TRA2015-69903-R Project (co-funded with ERDF funds), to the Spanish Ministry of Science, Innovation and Universities trough the project TRA2017-85853-C2-1-R, to the training grant FPU15 / 02990 of the Spanish Ministry of Education, Culture and Sports, to European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 688082 - SETA Project) and to the Full Professor Angel Ibeas, director of the Transport Systems Research Group.

Appendix A

A.1. Random parameter models

See Tables 7–9.

Table 7
Random parameter ordered Logit.

Attribute	Parameter	t-value
Non-random parameters		
Constant	−3.616	−5.35
Egress time	0.085	3.65
Vehicle cleanliness	0.221	3.45
Driver's kindness	0.057	3.75
Readability of map design	0.338	4.62
Noise	0.098	6.02
Service reliability/punctuality	0.576	7.17
Information on mobile app	0.071	2.93
Line coverage	0.412	5.51
Means for random parameters		
Use of hybrid buses	0.089	3.27
Access time to bus stop	0.340	5.67
Ease of transfer	0.311	6.79
Information at stops	0.333	4.76
Information on board	0.184	5.80
Comfort of the buses	0.137	3.98
Quality of bus stops	0.333	4.58
Information on the web page	0.213	5.14
Priority seats for PRM	0.139	4.45
Waiting time	0.301	4.97
Crowding level	0.332	5.18
In-vehicle travel time	0.211	6.59
Service frequency and timetables	0.566	7.57
Driving style	0.218	4.69
Calefaction/air conditioning	0.131	4.72
Scale parameters for dist. of random parameters		
Use of hybrid buses	0.063	9.43
Access time to bus stop	0.281	9.73
Ease of transfer	0.261	9.79
Information at stops	0.498	12.64
Information on board	0.201	10.74
Comfort of the buses	0.218	11.30
Quality of bus stops	0.150	6.31
Information on the web page	0.257	11.97
Priority seats for PRM	0.086	9.64
Waiting time	0.230	9.26
Crowding level	0.222	7.87
In-vehicle travel time	0.152	8.85
Service frequency and timetables	0.262	7.11
Driving style	0.152	13.54
CA	0.123	7.43
Threshold parameters		
Mu(01)	2.507	6.51
Mu(02)	6.835	14.27
Mu(03)	14.143	20.07
Log-likelihood		
AIC/N		
McFadden Pseudo R2		

Table 8
Mixed Logit model for Satisfaction BW data.

Attribute	Parameter	t-value
Means for random parameters		
Access time to bus stop	1.170	4.73
Waiting time	0.714	2.92
In-vehicle travel time	1.218	0.79
Egress time	1.054	8.18
Ease of transfer	0.910	6.94
Service frequency and timetables	0.307	2.2
Service reliability / punctuality	0.768	5.84
Line coverage	0.545	4.02

(*continued on next page*)

Table 8 (*continued*)

Attribute	Parameter	t-value
Information at stops	0.862	6.56
Information on the web page	0.493	3.94
Information on board	0.455	3.44
Crowding level	0.208	1.52
Calefaction / air conditioning	0.052	0.38
Priority seats for PRM	0.418	3.06
Comfort of the buses	0.670	5.12
Vehicle cleanliness	0.795	6.09
Driving style	0.072	0.53
Driver's kindness	0.478	3.61
Use of hybrid buses	1.328	10.06
Noise	0.071	0.53
Information on mobile app	0.587	4.24
Quality of bus stops	0.285	2.06
Readability of map design	0.114	0.82
Scale parameters for dist. Of random parameters		
Access time to bus stop	0.050	0.3
Waiting time	0.305	1.67
In-vehicle travel time	1.496	0.51
Egress time	0.131	1.43
Ease of transfer	0.074	0.82
Service frequency and timetables	0.050	0.5
Service reliability / punctuality	0.017	0.21
Line coverage	0.051	0.55
Information at stops	0.005	0.05
Information on the web page	0.024	0.29
Information on board	0.098	1.17
Crowding level	0.059	0.67
Calefaction / air conditioning	0.034	0.35
Priority seats for PRM	0.122	1.24
Comfort of the buses	0.100	1.06
Vehicle cleanliness	0.019	0.22
Driving style	0.058	0.54
Driver's kindness	0.017	0.2
Use of hybrid buses	0.066	0.73
Noise	0.047	0.49
Information on mobile app	0.091	0.91
Quality of bus stops	0.016	0.17
Readability of map design	0.031	0.3
Log-likelihood	−19.636	
AIC/N	16,239	

Table 9
Mixed Logit model for Importance BW data.

Attribute	Parameter	t-value
Non-random parameters		
Egress time	1.628	11.47
Service reliability / punctuality	2.337	17.43
Line coverage	2.235	16.57
Vehicle cleanliness	0.889	6.35
Driver's kindness	0.625	3.93
Noise	0.470	3.06
Information on mobile app	0.956	6.56
Readability of map design	1.113	7.59
Means for random parameters		
Access time to bus stop	1.076	4.07
Waiting time	1.808	8.72
In-vehicle travel time	2.810	2.29
Price / Fare	1.858	13.59
Ease of transfer	1.149	7.97
Service frequency and timetables	2.260	16.91
Information at stops	1.315	9.25

(*continued on next page*)

Table 9 (*continued*)

Attribute	Parameter	t-value
Information on board	0.245	1.48
Crowding level	1.489	10.78
Calefaction / air conditioning	0.425	2.76
Priority seats for PRM	1.619	11.47
Comfort of the buses	1.124	7.79
Driving style	1.471	10.19
Use of hybrid buses	1.081	7.76
Quality of bus stops	0.582	3.94
Scale parameters for dist. Of random parameters		
Access time to bus stop	0.302	1.7
Waiting time	0.163	1.25
In-vehicle travel time	1.092	0.47
Price / Fare	0.029	0.36
Ease of transfer	0.061	0.65
Service frequency and timetables	0.062	0.78
Information at stops	0.027	0.32
Information on board	0.147	1.14
Crowding level	0.142	1.9
Calefaction / air conditioning	0.024	0.22
Priority seats for PRM	0.014	0.17
Comfort of the buses	0.138	1.46
Driving style	0.035	0.37
Use of hybrid buses	0.044	0.52
Quality of bus stops	0.019	0.18
Log-likelihood	−19.636	
AIC/N	16,239	

A.2. Comparison of RP models

See Tables 10–11.

Table 10
correlation coefficients of means of RP models.

	Importance_BW	Satisfaction_BW	Ordered Logit	Satisfaction Rating
Importance_BW	1			
Satisfaction_BW	0.248	1		
Ordered Logit	0.417	0.060	1	
Satisfaction Rating	−0.102	0.806	0.007	1

Table 11
Correlation coefficients of scale parameters of RP models.

	Importance_BW	Satisfaction_BW	Ordered Logit	Satisfaction Rating
Importance_BW	1			
Satisfaction_BW	0.948	1		
Ordered Logit	0.203	0.053	1	
Satisfaction Rating	−0.105	−0.062	0.075	1

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tra.2019.10.012.

References

Abenoza, R.F., Cats, O., Susilo, Y.O., 2017. Travel satisfaction with public transport: Determinants, user classes, regional disparities and their evolution. Transportation Research Part A: Policy and Practice 95, 64–84. https://doi.org/10.1016/j.tra.2016.11.011.

Allen, J., Eboli, L., Forciniti, C., Mazzulla, G., Ortúzar, J. de D., 2019. The role of critical incidents and involvement in transit satisfaction and loyalty. Transport Policy 75, 57–69. https://doi.org/10.1016/j.tranpol.2019.01.005.

Allen, J., Eboli, L., Mazzulla, G., Ortúzar, J. de D., 2018a. Effect of critical incidents on public transport satisfaction and loyalty: an Ordinal Probit SEM-MIMIC approach. Transportation 1–37. https://doi.org/10.1007/s11116-018-9921-4.
Allen, J., Muñoz, J.C., Ortúzar, J. de D., 2018b. Modelling service-specific and global transit satisfaction under travel and user heterogeneity. Transportation Research Part A: Policy and Practice 113, 509–528. https://doi.org/10.1016/j.tra.2018.05.009.
Alonso, B., Barreda, R., Dell'Olio, L., Ibeas, A., 2018. Modelling user perception of taxi service quality. Transport Policy 63, 157–164. https://doi.org/10.1016/j.tranpol.2017.12.011.
Beck, M.J., Rose, J.M., 2016. The best of times and the worst of times: A new best-worst measure of attitudes toward public transport experiences. Transportation Research Part A: Policy and Practice 86, 108–123. https://doi.org/10.1016/j.tra.2016.02.002.
Beck, M.J., Rose, J.M., Greaves, S.P., 2017. I can't believe your attitude: a joint estimation of best worst attitudes and electric vehicle choice. Transportation 44, 753–772. https://doi.org/10.1007/s11116-016-9675-9.
Bordagaray, M., Dell'Olio, L., Ibeas, A., Cecín, P., 2014. Modelling user perception of bus transit quality considering user and service heterogeneity. Transportmetrica A: Transport Science 10, 705–721. https://doi.org/10.1080/23249935.2013.823579.
Brand, J., 1999. Development , implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.
Cao, J., Cao, X., 2017. Comparing importance-performance analysis and three-factor theory in assessing rider satisfaction with transit. Journal of Transport and Land Use 10. https://doi.org/10.5198/jtlu.2017.907.
Casella, G., George, E.I., Casella, G., George, E.I., 2016. Explaining the Gibbs Sampler Stable URL: http://www.jstor.org/stable/2685208 Linked references are available on JSTOR for this article: Explaining the Gibbs Sampler 3, 167–174.
de Oña, J., de Oña, R., 2015. Quality of service in public transport based on customer satisfaction surveys: A review and assessment of methodological approaches. Transportation Science 49, 605–622. https://doi.org/10.1287/trsc.2014.0544.
de Oña, J., de Oña, R., Eboli, L., Mazzulla, G., 2013. Perceived service quality in bus transit service: A structural equation approach. Transport Policy 29, 219–226. https://doi.org/10.1016/j.tranpol.2013.07.001.
de Oña, J., de Oña, R., Garrido, C., 2017. Extraction of attribute importance from satisfaction surveys with data mining techniques: a comparison between neural networks and decision trees. Transportation Letters 9, 39–48. https://doi.org/10.1080/19427867.2015.1136917.
Dell'Olio, L., Ibeas, A., Cecin, P., 2011. The quality of service desired by public transport users. Transport Policy 18, 217–227. https://doi.org/10.1016/j.tranpol.2010.08.005.
dell'Olio, L., Ibeas, A., Cecín, P., 2010. Modelling user perception of bus transit quality. Transport Policy 17, 388–397. https://doi.org/10.1016/j.tranpol.2010.04.006.
delÍOlio, L., Ibeas, A., de Oña, J., de Oña, R., 2018. Public Transportation Quality of Service. pp. 7–32.
Dyachenko, T., Reczek, R. W., Allenby, G. M., 2012. Models of Sequential Evaluation in Best-Worst Choice Tasks, SSRN. doi:10.2139/ssrn.2072496.
Eboli, L., Forciniti, C., Mazzulla, G., 2018. Spatial variation of the perceived transit service quality at rail stations. Transportation Research Part A: Policy and Practice 114, 67–83. https://doi.org/10.1016/j.tra.2018.01.032.
Eboli, L., Mazzula, G., 2009. A new customer satisfaction index for evaluating transit service quality. Journal of Public Transportation 12, 21–37. https://doi.org/10.5038/2375-0901.12.3.2.
Eboli, L., Mazzulla, G., 2015. Relationships between rail passengers satisfaction and service quality: a framework for identifying key service factors. Public Transport 7, 185–201. https://doi.org/10.1007/s12469-014-0096-x.
Eboli, L., Mazzulla, G., 2011. A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view. Transport Policy 18, 172–181. https://doi.org/10.1016/j.tranpol.2010.07.007.
Echaniz, E., Dell'Olio, L., Ibeas, Á., 2018. Modelling perceived quality for urban public transport systems using weighted variables and random parameters. Transport Policy 67, 31–39. https://doi.org/10.1016/j.tranpol.2017.05.006.
Echaniz, E., Ho, C., Rodriguez, A., Dell'Olio, L., 2019. Modelling user satisfaction in public transport systems considering missing information. Transportation. https://doi.org/10.1007/s11116-019-09996-4.
Efthymiou, D., Antoniou, C., Tyrinopoulos, Y., Skaltsogianni, E., 2018. Factors affecting bus users' satisfaction in times of economic crisis. Transportation Research Part A: Policy and Practice 114, 3–12. https://doi.org/10.1016/j.tra.2017.10.002.
Gilks, W.R. (Wally R.., Richardson, S. (Sylvia), Spiegelhalter, D.J., 1996. Markov chain Monte Carlo in practice. Chapman & Hall.
Gonzalo-Orden, H., dell'Olio, L., Ibeas, A., Rojo, M., 2011. Modelling gender perception of quality in interurban bus services. Proceedings of the ICE - Transport 164, 43–53. https://doi.org/10.1680/tran.9.00031.
Greene, W.H., 2016. Nlogit 6 guide.
Guirao, B., Eugenia López, M., Comendador, J., 2015. New QR Survey Methodologies to Analyze User Perception of Service Quality in Public Transport: The Experience of Madrid. Journal of Public Transportation 18, 1–5.
Guirao, B., García-Pastor, A., López-Lambas, M.E.M.E., Garc??a-Pastor, A., L??pez-Lambas, M.E., García-Pastor, A., López-Lambas, M.E.M.E., 2016. The importance of service quality attributes in public transportation: Narrowing the gap between scientific research and practitioners' needs. Transport Policy 49, 68–77. doi: 10.1016/j.tranpol.2016.04.003.
Hensher, D.A., Stopher, P., Bullock, P., 2003. Service quality - developing a service quality index in the provision of commercial bus contracts. Transportation Research Part A: Policy and Practice 37, 499–517. https://doi.org/10.1016/S0965-8564(02)00075-7.
Ho, C.Q., Hensher, D.A., 2017. Application of irrelevance of state-wise dominated alternatives (ISDA) for identifying candidate processing strategies and behavioural choice rules adopted in best–worst stated preference studies. Journal of Choice Modelling 25, 40–49. https://doi.org/10.1016/J.JOCM.2017.01.002.
Horton, N.J., Lipsitz, S.R., Horton, N.J., Lipsitz, S.R., 2016. Multiple Imputation in Practice : Comparison of Software Packages for Regression Models with Missing Variables Statistical Computing Software Reviews Multiple Imputation in Practice : Comparison of Software Packages for Regression Models With Missing Vari 55, 244–254.
Ibeas, A., dell'Olio, L., Montequín, R.B., 2011. Citizen involvement in promoting sustainable mobility. Journal of Transport Geography 19, 475–487. https://doi.org/10.1016/j.jtrangeo.2010.01.005.
Lipovetsky, S., Conklin, M., 2014. Best-Worst Scaling in analytical closed-form solution. Journal of Choice Modelling 10, 60–68. https://doi.org/10.1016/j.jocm.2014.02.001.
Louviere, J.J., Flynn, T.N., Marley, A.A.J., 2015. Best-worst scaling: Theory, methods and applications, Best-Worst Scaling: Theory, Methods and Applications. doi: 10.1017/CBO9781107337855.
Marley, A.A.J., Islam, T., Hawkins, G.E., 2016. A formal and empirical comparison of two score measures for best–worst scaling. Journal of Choice Modelling 21, 15–24. https://doi.org/10.1016/J.JOCM.2016.03.002.
Marley, A.A.J., Louviere, J.J., 2005. Some probabilistic models of best, worst, and best-worst choices. Journal of Mathematical Psychology 49, 464–480. https://doi.org/10.1016/j.jmp.2005.05.003.
Marley, A.A.J., Pihlens, D., 2012. Models of best–worst choice and ranking among multiattribute options (profiles). Journal of Mathematical Psychology 56, 24–34. https://doi.org/10.1016/J.JMP.2011.09.001.
Martilla, J.A., James, J.C., 1977. Importance-Performance Analysis. Journal of Marketing 41, 77. https://doi.org/10.2307/1250495.
Mouwen, A., 2015. Drivers of customer satisfaction with public transport services. Transportation Research Part A: Policy and Practice 78, 1–20. https://doi.org/10.1016/j.tra.2015.05.005.
Mulley, C., Hensher, D.A., Rose, J., 2014. Do preferences for BRT and LRT vary across geographical jurisdictions? A comparative assessment of six Australian capital cities. Case Studies on Transport Policy 2, 1–9. https://doi.org/10.1016/j.cstp.2013.11.001.
Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P., 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27, 85–95.
Román, C., Martín, J.C., Espino, R., 2014. Using Stated Preferences to Analyze the Service Quality of Public Transport. International Journal of Sustainable Transportation 8, 28–46. https://doi.org/10.1080/15568318.2012.758460.

Rose, J.M., Hensher, D.A., 2018. User satisfaction with taxi and limousine services in the Melbourne metropolitan area. Journal of Transport Geography 70, 234–245. https://doi.org/10.1016/j.jtrangeo.2018.06.017.

Rubin, D. B., 1978. Multiple imputations in sample surveys - A phenomeno- logical Bayesian approach to nonresponse. In Proceedings of the Survey Research Methods Section of the American Statistical Association (pp. 20–28).

Rubin, D.B., 1977. Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys Formalizing Sub jective Notions About the Effect of Nonrespondents in Sample Surveys. Source Journal of the American Statistical Association 72144202, 538–543.

Sever, I., 2015. Importance-performance analysis: A valid management tool? Tourism Management 48, 43–53. https://doi.org/10.1016/J.TOURMAN.2014.10.022.

Tyrinopoulos, Y., Antoniou, C., 2008. Public transit user satisfaction: Variability and policy implications. Transport Policy 15, 260–272. https://doi.org/10.1016/j.tranpol.2008.06.002.

van Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research 16, 219–242. https://doi.org/10.1177/0962280206074463.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., Rubin, D. B., 2006. Fully conditional specification in multivariate imputation 76, pp. 1049–1064. doi: 10.1080/10629360600810434.

Weinstein, A., 2000. Customer satisfaction among transit riders: How customers rank the relative importance of various service attributes. Transportation Research Record: Journal of the Transportation Research Board 1735, 123–132. https://doi.org/10.3141/1735-15.