**Article:**

Song, F, Hess, S and Dekker, T (2021) A joint model for stated choice and best-worst scaling data using latent attribute importance: application to rail-air intermodality. Transportmetrica A: Transport Science, 17 (4). pp. 411-438. ISSN: 2324-9935

https://doi.org/10.1080/23249935.2020.1779384

# A joint model for stated choice and best-worst scaling data using latent attribute importance: application to rail-air intermodality

Fangqing Song[a,b], Stephane Hess[a,b] and Thijs Dekker[a,b]

[a]Institute for Transport Studies, University of Leeds, UK;
[b]Choice Modelling Centre, University of Leeds, UK

**ARTICLE HISTORY**
Compiled May 22, 2020

**ABSTRACT**
This paper looks at modelling choices in the presence of a new mode of transport, where there is need to understand the sensitivities to a number of new attributes. Stated choice (SC) data and two types of Best-worst scaling (BWS) data (i.e. case 1 and case 2) are collected from the same respondents. We mix survey methods rather than using a longer SC survey to better understand choice behaviour whilst reducing the boredom caused by one very long set of SC choices. Although BWS data has been increasingly collected alongside stated choice (SC) data, little is known about the relationships between BWS responses and SC responses at the level of individual respondents. Also, little effort has been made to jointly exploit the behavioural information from BWS data and SC data to improve the understanding of choices. This paper proposes a joint model which links the BWS and SC data through the notion of latent *attribute importance*. The modelling results show that people perceive *attribute importance* in a relatively consistent way across different survey methods, i.e. a person who perceives higher importance from an attribute is likely to show stronger sensitivity to that attribute in SC tasks, give more weight to the same attribute in BWS1 tasks and exhibit a wider gaps in terms of attractiveness between levels for the same attribute - in comparison with other individuals. This consistency shows that the additional behavioural information can be gained by using a joint model estimated on BWS1 and BWS2 data alongside more traditional SC data, helping us to improve the explanation of the choices and the role of the attributes. Our results however do not find a one-to-one relationship between different survey methods and analysts thus need to be mindful that there remain some differences in how attributes are evaluated between SC, BWS1 and BWS2 surveys.

**KEYWORDS**
Stated choice, Best-worst scaling, Attribute importance, MaxDiff model, Integrated Choice and Latent Variable model

CONTACT Stephane Hess. Email: S.Hess@leeds.ac.uk

## 1. Introduction

Many new travel modes have emerged in recent years. Studies aimed at understanding individuals' choice behaviour and the travel demand for novel alternatives have predominantly relied on stated-choice (SC) data, where a respondent chooses his/her most preferred alternative in each hypothetical scenario. A new travel mode is usually characterised with some new attributes which individuals are not familiar with. Therefore, a key role of the surveys is to gain more information on how these new attributes are valued by respondents. These attributes are often not continuous in nature and the reliable estimation of their impact can thus require substantial amounts of data. However, increasing the number of tasks of a SC survey might lead to respondents feeling greater boredom to process a repeated same type of choice tasks. Thus, it can be useful to gain additional behavioural information through other types of preference elicitation methods to help us better understand how people make choices in the context of new modes and the role that these new attributes play. This combination of data sources can be helpful to improve the robustness of policy recommendations. This can especially be the case when the number of tasks that can be used in an SC experiment is limited due to the increasing boredom brought on by a longer set of repeated SC tasks. Moreover, respondents may experience fatigue in a SC survey where many attributes are presented all at the same time (Pullman, Dodson, and Moore 1999; Carlsson 2003; Collins, Bliemer, and Rose 2014).

Recently, a limited number of travel behaviour studies have adopted best-worst scaling (BWS) approaches as alternative preference elicitation methods (e.g. Dumont, Giergiczny, and Hess 2015; Hensher, Mulley, and Rose 2015; Beck and Rose 2016; Beck, Rose, and Greaves 2017). The BWS approaches originate in marketing and the majority of its applications can be found in the marketing and health literature. In BWS, respondents are asked to in each task select the best and the worst option. Different formats of this exist. BWS Case 1 surveys ask respondents to identify, in each choice screen, the most and the least important attribute per se without a focus on the actual levels (e.g. Finn and Louviere 1992; Auger, Devinney, and Louviere 2007; Marti 2012). BWS Case 2 surveys ask respondents to identify the most and the least important attribute level (e.g. Coast et al. 2006; Dyachenko, Reczek, and Allenby 2014). While BWS Case 1 measures the relative weight of attributes, BWS Case 2 measures the relative attractiveness of attribute levels across different attributes.[1] Like SC surveys, BWS Case 3 surveys also look at comparisons amongst different alternatives, each described by a combination of attribute levels; but BWS Case 3 surveys require respondents to identify both the most and the least preferred alternative in each choice

[1] In this presented paper, we use *weight* to describe the influence of an attribute in decision making in BWS Case 1 tasks and use *attractiveness* to describe the influence of an attribute level in decision making in BWS Case 2 tasks. Greater *weight* of an attribute or *attractiveness* of an attribute level means higher probability of this attribute or attribute level being chosen as the best and lower probability of it being chosen as the worst.

occasion. Comparisons between SC and BWS case 3 data can be found in the work of Giergiczny et al. (2017) and Petrolia, Interis, and Hwang (2018).

This research is conducted in the context where a new travel mode, i.e. high-speed rail(HSR)-air intermodality, is introduced. Since our interest is in predicting choices (i.e. first preferences only), we adopt a traditional SC survey as it allows us to analyse how respondents make trade-offs between attributes and forecast travel demand within multi-alternative settings. A BWS Case 3 survey is not adopted for this purpose as it combines both the best and the worst where existing studies show diverging views on how consistent people are in choosing the best and the worst. Some found differences in both utility parameters and scales between the two stages (Rose 2014; Giergiczny et al. 2017), notwithstanding contrary findings in Hawkins, Islam, and Marley (2018) that suggested that the same utility parameters drive individuals' best and worst choices despite a scale difference between best choices and worst choices. In addition to the SC survey, BWS Case 1 and BWS Case 2 surveys are used as these two methods can reflect how individuals are influenced by different attributes in relatively more direct manners in single-alternative settings. As such, BWS Case 1 and BWS Case 2 data serves as additional behavioural information to help in better explaining the role of specific attributes in these choice decisions.[2]

This paper aims at exploring approaches to synthesise SC, BWS Case 1 and Case 2 data within a same modelling framework to understand their relationships at the level of individual respondents and to improve the explanation of choices with the help of the supplementary information obtained from BWS Case 1 and Case 2 data. A key question in achieving this target, which has not been addressed in the literature, is whether the extent to which respondents weight attributes in a BWS Case 1 survey and rank attribute levels in a BWS Case 2 survey is consistent with how those same attributes and levels influence the choices in a SC survey. A higher level of correspondence between the different data sources would imply greater exploitation of the auxiliary BWS Case 1 and Case 2 data in enhancing the explanation of stated choices and building a more robust evidence base for policy recommendations.

The majority of studies comparing SC data and BWS Case 1 and (or) Case 2 data have been conducted at the sample level (e.g. Louviere and Islam 2008; Potoglou et al. 2011). Only Balbontin, Ortúzar, and Swait (2015) and Beck, Rose, and Greaves (2017) have jointly analysed SC and BWS Case 2 data. However, there are some remaining limitations associated with these two joint estimation studies. The former lacks flexibility in model specifications as it assumes the impact of an attribute level in the SC tasks to be equal (or a function of) to the impact of the same attribute level revealed in the BWS Case 2 data. The latter directly incorporates the average

[2] BWS approaches outweigh rating or ranking methods as BWS can take advantage of respondents' tendency of responding more consistently and accurately to extreme options on an underlying scale from a relatively small choice set(Marley and Louviere 2005). Thus conventional rating or ranking tasks are not used to help explain choices in our study.

impact over different attribute levels from BWS Case 2 data to help explain choices in SC data and thereby exposes itself to potential endogeneity biases. Meanwhile, joint analyses of SC data with BWS Case 1 data have not yet been explored.[3]

In this paper, we put forward a flexible approach to jointly estimate SC, BWS Case 1 and BWS Case 2 data at the individual level while overcoming the shortcomings in the literature. This approach is based on the assumption that responses to BWS Case 1, BWS Case 2 and SC tasks are all driven by a common underlying factor of perceived *attribute importance*. We develop an Integrated Choice and Latent Variable (ICLV) model (Ben-Akiva et al. 2002) where each attribute is associated with a specific latent variable of *attribute importance*. The notion of *attribute importance* has previously been put forward to challenge the decision heuristic of attribute non-attendance (Hensher, Rose, and Greene 2005; Hensher 2006; Hensher and Rose 2009), arguing that some people actually perceive reduced importance for an attribute in making stated choices rather than completely ignoring it even if the respondents stated that they did not take the associated attribute into account (Hess and Hensher 2010; Campbell, Hensher, and Scarpa 2011; Hess et al. 2013). Our work adopts a similar strategy as Hess and Hensher (2013), who use latent *attribute importance* to simultaneously explain the responses to SC tasks and the responses to selected indicators, including binary stated attribute attendance and stated attribute rankings. In our proposed model, the indicators are replaced by BWS Case 1 and Case 2 data.

We apply the proposed model in the context of a new HSR-air intermodal service in China. This new service facilitates people's long-distance travel by allowing passengers to jointly take HSR trains and flights to make a journey without the hassle of purchasing train tickets and flights separately. As expected, we find a certain degree of correspondence among the behaviour in the stated choice scenarios, BWS Case 1 exercises and BWS Case 2 exercises. That is, for a given attribute, people who perceive stronger importance of an attribute derive higher marginal utility from that attribute in SC tasks, attach higher weight to that attribute in BWS1 tasks, and are more sensitive to changes in level values of that attribute in BWS2 tasks - in comparison with other people. This correlation suggests that the supplementary BWS1 and BWS2 tasks can indeed bring about desired additional information and help better explain the role of attributes. There is, however, not a one-to-one relationship between the different survey methods. This implies that researchers, while being keen to explore the additional insights provided by BWS data, should not treat SC and BWS survey methods as equivalent and interchangeable.

The remainder of this paper is organised as follows. Section 2 explains the methodology of the joint model. The survey design and the data is described in section 3. The

[3]BWS Case 1 and SC data is often collected at different moments of the survey design and collection process. Outcomes of the former are for example regularly used to determine which attributes from a larger pool of attributes need to be included in the SC experiment.

case study is analysed in section 4, which is followed by a conclusion section.

## 2. Methodology

In this section, we look at the individual components of our model framework before discussing estimation results. For the sake of brevity, we use “BWS1” and “BWS2” to represent “BWS Case 1” and “BWS Case 2” respectively.

### 2.1. *Model framework*

As mentioned in the Introduction, our model is developed based on the assumption of correlation between SC responses and BWS1/2 responses. Latent variables are introduced to capture the correlation and to simultaneously explain different types of responses within a single ICLV framework. We follow the adoption of the notion *attribute importance* from Hess and Hensher (2013) to represent latent variables for each attribute as SC, BWS1 and BWS2 surveys all reveal people's preferences towards various attributes in the decision-making process.[4]

Fig. 1 illustrates our joint modelling framework, where items in rectangulars are observable to researchers while items in ellipses are unobserved. Brief descriptors of each notation used in section 2, including those appeared in Fig. 1, are shown in the Appendix. The model has three components, explaining the SC responses $y$, BWS1 responses $(b, w)_{|1}$ and BWS2 responses $(b, w)_{|2}$ respectively. The latter two form the measurement model components. All three components are influenced and connected by the attribute-specific latent variable of *attribute importance*. As such, we do not impose restrictions on how an attribute (or attribute level) is evaluated between BWS1/2 data and SC data as in the work by Balbontin, Ortúzar, and Swait (2015). We also do not directly feed the BWS1 and BWS2 responses as explanatory variables into the choice model component as Beck, Rose, and Greaves (2017) did. Thereby, the proposed model has greater flexibility in recovering the correlations between BWS and SC responses,and data collected through different methods can be synthesised without the risk of introducing endogeneity bias or measurement error.

More precisely, the attribute-specific latent variables of *attribute importance* are used as explanatory variables for each elicitation procedure. For each specific attribute, we assume that the corresponding *attribute importance* scales the marginal utility of that attribute in the SC component, hence influencing the utilities of alternatives in the utility functions which are also affected by some socioeconomic characteristics. Meanwhile, the latent *attribute importance* also determines the same attribute's weight

[4]Please refer to the definition of *weight* and *attractiveness* in footnote 1. It also needs to be noted that our definition of *attribute importance* is not equivalent to the *importance* defined by Marley, Flynn, and Louviere (2008), and we do not have the same identifiability problem as discussed in that paper as we are not trying to separate the impact of an attribute and a specific level on that attribute in BWS2 tasks.

in the BWS1 component as well as the attractiveness of attribute levels of the same attribute in the BWS2 component. Different coefficients are specified to capture the different impact of a same latent *attribute importance* in different methods. In what follows, we discuss how each component is constructed and the role of latent *attribute importance* in detail.

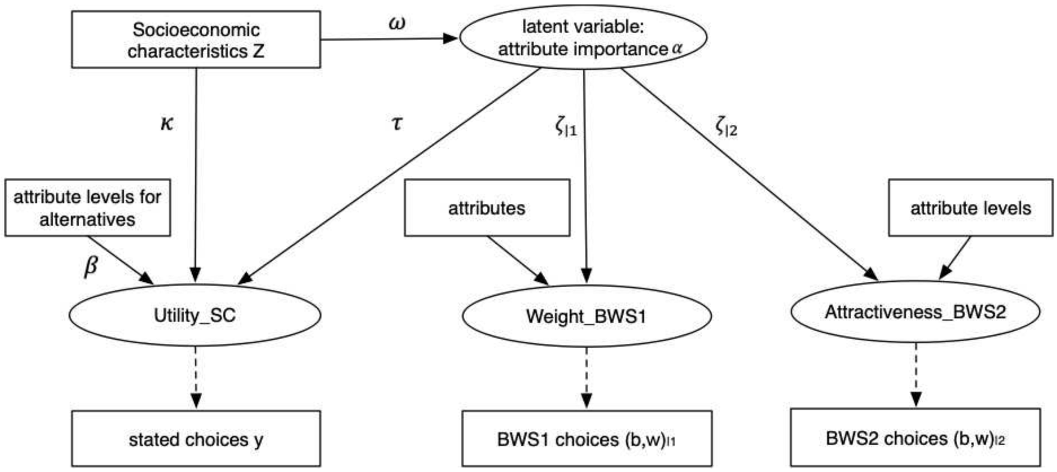


Figure 1.: Framework of the joint model.

## 2.2. *Structural equations for latent variables*

We denote the attribute-specific latent variables of *attribute importance*, as perceived by respondent $n$, by the vector $\alpha_n = (\alpha_{n1}, \ldots, \alpha_{nK})'$, where $K$ describes the total number of attributes. Selected socio-demographic characteristics $Z_n$ are used to explain the latent variables in the structural equations:

$$\alpha_{nk} = \omega_k' Z_n + \eta_{nk}, \quad (k = (1, \cdots, K)), \tag{1}$$

where $\eta_{nk}$ is a standard Normal error term and where the estimated vector of parameters $\omega_k$ measures the impact of the socio-demographic characteristics on the latent variable. Note that $Z_n$ is centred on 0, such that the latent variable $\alpha_{nk}$ has a mean of 0.

## 2.3. *Stated choice model component*

The model is constructed under the Random Utility Maximisation (RUM) theory, where it is presumed that a decision-maker can derive some utility from choosing a particular alternative and that the probability of choosing an alternative increases with its utility.

Let $U_{int}$ in Eq. 2 represent the utility of alternative $i$ for respondent $n$ in stated choice task $t$. $U_{int}$ consists of a deterministic portion $V_{int}$ (i.e. systematic utility), and an unobserved error term $\varepsilon_{int}$ which is independently and identically distributed (IID) extreme value type I.

$$U_{int} = V_{int} + \varepsilon_{int} = \delta_i + \beta_n' x_{int} + \varepsilon_{int}. \tag{2}$$

The term $\delta_i$ is an estimated alternative-specific constant (ASC) while $x_{int} = (x_{int1}, \cdots, x_{intK})'$ is a vector of explanatory variables representing the $K$ attributes of alternative $i$ as shown to respondent $n$ in SC task $t$, where the estimated vector $\beta_n = (\beta_{n1}, \cdots, \beta_{nK})'$ captures the marginal utilities of these attributes. Hence, it is assumed that each attribute contributes to the utility of an alternative in an additive manner, and that the marginal utility for each attribute is kept generic across alternatives.

Marginal utility varies across respondents due to the role of the latent *attribute importance*, as well as additional observed and unobserved preference heterogeneity that is independent of the latent variable. For an attribute where we assume a positive marginal utility, we specify $\beta_{nk}$ such that:

$$\beta_{nk} = e^{\tau_k \alpha_{nk}} \cdot e^{\kappa_k Z_n} \cdot e^{\mu_{\ln\beta_k} + \sigma_{\ln\beta_k} \cdot \xi_{nk}}, \tag{3}$$

where, for an attribute with an expected negative marginal utility, we instead work with the negative exponential such that:

$$\beta_{nk} = -e^{\tau_k \alpha_{nk}} \cdot e^{\kappa_k Z_n} \cdot e^{\mu_{\ln(-\beta_k)} + \sigma_{\ln(-\beta_k)} \cdot \xi_{nk}}. \tag{4}$$

Latent *attribute importance* is accommodated in an exponential form to act as a positive scalar on marginal utility where $\tau_k$ captures the degree of scaling (Hess and Hensher 2013). To avoid overstating the role of latent *attribute importance* in explaining heterogeneity in the SC data (Vij and Walker 2016), we let the socio-demographics $Z_n$ which explain the latent variable $\alpha_{nk}$ in the structural equations also directly enter the marginal utility, where the vector $\kappa_k$ measures the direct impacts from socio-demographics $Z_n$ on the scaling of marginal utility. Additional random heterogeneity that is not linked to the latent variable is accommodated by specifying the underlying parameter, net of the influence of socio-demographics and the latent variable, to follow a Lognormal distribution. We then have that $\mu_{\ln\beta_k}$ and $\sigma_{\ln\beta_k}$ (or $\mu_{\ln(-\beta_k)}$ and $\sigma_{\ln(-\beta_k)}$ if we work with a negative exponential) denote the mean and standard deviation of the underlying Normal distribution, where $\xi_{nk}$ follows a standard Normal distribution across respondents for attribute $k$. It can be observed that as

$e^{\tau_k \alpha_{nk}}$ itself follows a Lognormal distribution, $\beta_{nk}$ does too as it is formed by a product of Lognormals.

The probability of alternative $s$ being chosen out of $I$ alternatives by respondent $n$ in SC task $t$ is then written as:

$$P(y_{nt} = s) = \frac{e^{\delta_s + \sum_{k=1}^{K} \beta_{nk} x_{sntk}}}{\sum_{i=1}^{I} e^{\delta_i + \sum_{k=1}^{K} \beta_{nk} x_{intk}}}, \tag{5}$$

where this is dependent on a specific realisation of the vector of random coefficients.

### 2.4. *Measurement model components*

In explaining BWS1 and BWS2 data, we develop models based on the MaxDiff model (Marley and Louviere 2005; Marley, Flynn, and Louviere 2008), attempting to explain the choice for the observed pair of best and worst attributes $(b, w)_{|1}$, and attribute levels $(b, w)_{|2}$, respectively. MaxDiff models explain the choice of the combination of attributes or attribute levels with the largest difference in "utility" between them. In the remainder of this paper, we use "utility" to refer to the weight of an attribute in the BWS1 component and the attractiveness of an attribute level in the BWS2 component, for the sake of brevity.[5]

Let $B_{qnm|c}$ denote the "utility" of $q$ for respondent $n$ as shown in BWS task $m$ and BWS type $c$, where $c = 1$ stands for BWS1 and $c = 2$ for BWS2. We thus define:

$$BW_{(q,j)nm|c} = B_{qnm|c} + W_{jnm|c} + \nu_{qjnm|c}, \tag{6}$$

where $B_{qnm|c}$ and $W_{jnm|c}$ give the "utility" of the two attributes or attribute levels that would be used to create the combination $(q, j)$ while $\nu_{qjnm|c}$ denotes a standard extreme value type I error term operating at the level of the attribute (level) pairs allowing us to operate within the Multinomial Logit (MNL) framework when deriving the probability of a given pair being the one with the largest difference in "utility". Rather than simply assuming symmetry between the "utilities" for the best and the worst levels, we set:

$$W_{jnm|c} = -\lambda_{j|c} B_{jnm|c}, \tag{7}$$

thus accounting for scale difference between the "best" and the "worst" stage and allowing this difference to be attribute-specific, while still assuming that the driving factors of making an attribute (level) important/attractive or unimportant/unattractive are the same across the two stages. Hence this specification is different from the original

[5] The quoted term "utility" is used for precision as *utility* by definition can only be derived from an alternative (McFadden et al. 1973; McFadden 2001), rather than from a single attribute or attribute level.

MaxDiff model proposed by (Marley and Louviere 2005; Marley, Flynn, and Louviere 2008), where scale parameters were not included (i.e. $\lambda_{j|c} = 1$). We thereby refer to our models for the BWS1/2 data as *MaxDiff models with scale difference.*

Due to the experimental design, the choice set varies over respondents and tasks, and this thus affects what is possible for a respondent to select as the combination of best and worst attributes or attribute levels in a given scenario. We use $\mathbb{D}_{nm|c}$ to define the set containing all the available items presented to respondent $n$ in BWS task $m$ and type of BWS data $c$. The items in $\mathbb{D}_{nm|c}$ allow the formation of the set $\mathbb{S}_{nm|c}$ containing all the possible best-worst pairs of the available attributes or attribute levels, respectively. Similar to other MNL models with a RUM assumption, the best-worst choice probabilities of respondent $n$ selecting $h$ as the best and $r$ as the worst $(h, r \in \mathbb{D}_{nm|c}, r \neq h, (h, r) \in \mathbb{S}_{nm|c})$ in BWS task $m$ can then be written as:

$$P\left((b, w)_{nm|c} = (h, r)\right) = \frac{e^{B_{hnm|c}+W_{rnm|c}}}{\sum_{(q,j)\in\mathbb{S}_{nm|c}}\left(e^{B_{qnm|c}+W_{jnm|c}}\right)}, \tag{8}$$

making use of the appropriate combinations of Eqs. 9 - 11 discussed in what follows.

### *2.4.1. BWS1 data*

In the BWS1 setting, we work with attributes rather than attribute levels. The "utility" function is specified to represent the weight placed on an attribute $k$ by respondent $n$ in task $m$ in decision-making. Thus we have a single "utility" for a given attribute $k$ to be "best" attribute, which is given by:[6]

$$B_{knm|1} = \delta_{k|1} + \zeta_{k|1}\alpha_{nk}, \tag{9}$$

where this is generic across BWS1 tasks as the attribute levels are not used. In Eq. 9, we have a constant $\delta_{k|1}$ and a sensitivity $\zeta_{k|1}$ with respect to the latent variable, where these two parameters are to be estimated. Since $\alpha_{nk}$ is centred on 0, $\delta_{k|1}$ captures the mean weight of attribute $k$ in the BWS1 data, while $\zeta_{k|1}$ captures the variation in the weight of the attribute in the sample due to latent attribute importance. Respondents who perceive a higher importance to an attribute are expected to care more about that attribute in the BWS1 data.

For normalisation purpose, one attribute in the MaxDiff model with scale difference

[6] In an ICLV model, it is common practice to use the latent variable solely to capture heterogeneity in the measurement component, and only a limited number of studies have also directly included additional randomness irrelevant from the latent variable in the measurement model. We have tried to estimate models with such direct random component in the measurement model for the BWS1 data. However, log-likelihood ratio test suggests accounting for such randomness cannot bring about significant improvement in fit or help better explain choices in our case. The interpretation of the estimation results are nevertheless quite similar to the old model, indicating that our findings about the correlation among different survey methods are relatively consistent across different model specifications. This also applies to the specification for BWS2 data in Eqs. 10 and 11.

for BWS1 data needs to be selected as the base by fixing the associated parameters to 0.

### 2.4.2. BWS2 data

In the BWS2 data, we work with multiple levels across attributes. The BWS2 "utility" function describes the attractiveness of an attribute level (or value) $k$ perceived by respondent $n$ in task $m$. The specification for a given attribute level $k$ now depends on whether this attribute is treated as continuous or categorical. We explicitly here do not allow for scenarios in which multiple values for the same attribute are shown on one screen, i.e. only allowing for screens where each element is from a different attribute.

Let us define $x_{knm|2}$ to be the value of continuous variable $k$ as shown in BWS2 task $m$ for respondent $n$. We then define $B_{knm|2}$ to be equal to:

$$B_{knm|2} = \delta_{k|2} + \gamma_{k|2} \cdot e^{\zeta_{k|2}\alpha_{nk}} x_{knm|2}. \tag{10}$$

Here, we assume that the attractiveness of a level depends in a linear fashion on the actually presented value $x_{knm|2}$, $\delta_{k|2}$ captures the constant associated with attribute $k$ and $\gamma_{k|2}$ captures the baseline marginal attractiveness of the attribute level on $B_{qnm|2}$. This marginal attractiveness is then affected by the latent variable, where $\zeta_{k|2}$ scales the level spacing based on latent attribute importance.

The treatment is different if attribute $k$ is a categorical variable. In that case, a specific level will apply. Let us assume that attribute $k$ takes $L_k$ possible values in a survey. We would then have:

$$B_{knm|2} = \phi_{k_1|2} \left(x_{knm|2} == 1\right) + \sum_{l=2}^{L_k} \phi_{k_l|2} \left(e^{\zeta_{k|2}\alpha_{nk}}\right) \left(x_{knm|2} == l\right). \tag{11}$$

In this specification, we have a sum over all the possible levels that could apply for attribute $k$, where only one of these will apply in a given BWS2 scenario, and where the bracket $\left(x_{knm|2} == l\right)$ will be equal to 1 for that specific level. We now estimate the baseline attractiveness of each level for the categorical attribute through $\phi_{k_l|2}$. The baseline attractiveness parameter $\phi_{k_l|2}$ is then further re-scaled by the corresponding latent attribute importance through $\zeta_{k|2}$, where this impact of the latent variable is attribute rather than attribute-level specific. We do not scale the base level (i.e. $l = 1$) to avoid the situation where an individual with higher *attribute importance* derives higher attractiveness from the base level of attribute $k$ than other individuals. Under the current specification, respondents with higher attribute importance then exhibit a wider gap in terms of attractiveness between a higher level and the lowest (base) level for that attribute than others do.

For normalisation purpose, one attribute level across all attributes in the MaxDiff model with scale difference for BWS2 data needs to be selected as the base by fixing the associated parameters to 0.

### 2.5. *Log-likelihood*

The unconditional probability of observing the sequence of stated choices $y_n$ and best-worst responses $(b,w)_n$ can be expressed as the integral of the multiplication of the conditional stated choice probabilities and the conditional best-worst choice probabilities over the distribution of $\eta_n$, the random component of the latent variables $\alpha_n$, and over the distribution of $\xi_n$, the random component of the unobserved preference heterogeneity irrelevant from $\alpha_n$, such that the log-likelihood is given by:

$$\begin{aligned} &LL(y,(b,w)) = \\ &\sum_{n=1}^{N} ln \int_{\xi_n}\int_{\eta_n}\left(\prod_{t=1}^{T_n} P\left(y_{nt}\mid\beta_n\right)\prod_{m|1=1}^{M_{n|1}} P\left((b,w)_{nm|1}\mid\alpha_n\right)\prod_{m|2=1}^{M_{n|2}} P\left((b,w)_{nm|2}\mid\alpha_n\right)\right), \\ &f\left(\eta_n\right) g\left(\xi_n\right) \mathrm{d}\eta_n \mathrm{d}\xi_n \end{aligned} \tag{12}$$

where $T_n$, $M_{n|1}$ and $M_{n|2}$ give the total numbers of the SC tasks, the BWS1 tasks, and the BWS2 tasks shown to respondent $n$. Meanwhile, choice observations $y_{nt}$, $(b,w)_{nm|1}$, $(b,w)_{nm|2}$ refer to the chosen alternative in a SC task, the chosen best-worst pair of attributes in a BWS1 task, and the best-worst pair of attribute levels selected in a BWS2 task, respectively. Since the resulting $LL$ does not have closed-form expression, the value of the log-likelihood needs to be approximated through simulation (Train 2009).

### 2.6. *Hypothesis*

A hypothesis is put forward with respect to the correlations among stated choices, BWS1 responses and BWS2 responses as well as the role of latent *attribute importance* in the joint model. Providing that a higher value of the latent variable is associated with stronger *attribute importance*, we expect the signs of the impact factors of *attribute importance* in the choice model and measurement models (i.e. $\tau$,$\zeta_{|1}$,$\zeta_{|2}$) to all be positive. That is, respondents who perceive higher importance from an attribute would have a higher probability to:

- be more sensitive (i.e. higher marginal utility) to the attribute in SC tasks;

- give more weight to the same attribute per se in BWS1 tasks;
- experience a wider gap in terms of attractiveness between a higher level and the lowest level (i.e. higher marginal attractiveness) for the attribute concerned in BWS2 tasks.

Of course, the same result also applies if all signs are negative, i.e. a higher latent variable leads to lower sensitivities in SC, lower weights in BWS1 and narrower attractiveness gaps in BWS2. In that case, the latent variable would be interpreted as *reduced attribute importance*. Opposite signs for the different effects or insignificance indicate a lack of consistency for the associated attribute across datasets. If fixing all the impact factors to 0, the joint ICLV model would be equivalent in specification to a model which pools all the three datasets but ignores any correlations in between. In this sense, our model can identify to what extent the choices made and the role of attributes played are consistent across different types of tasks, and explore whether the behavioural information contained in BWS1 and BWS2 data could help improve the understanding of SC data.

It is worth noting that the latent variables of *attribute importance* are not used to show the influence on an attribute in comparison with other attributes, but instead to explain part of the variation across individuals. That is, if the hypothesis can be confirmed, ceteris paribus, a higher value of the latent attribute importance $\alpha_{nk}$ would mean individual $n$ is relatively more strongly influenced by attribute $k$ in different tasks than other individuals, rather than indicating perceiving more importance from attribute $k$ than from other attributes.

## 3. Case study: Survey and data

### 3.1. *Survey background*

Our research is conducted in the context of HSR (high-speed rail)-air intermodality in China. This integrated HSR-air service has been put into practice since 2011 in Shanghai with an aim to enhance the connectivity of Shanghai and its non-airport catchment area by enabling passengers to jointly travel by HSR and air on a single trip with a convenient and even seamless transfer between the two different modes and without the need of purchasing HSR and flight tickets separately.

Since collecting data from real passengers at an airport terminal is very difficult,[7] we tried to gain more behavioural and preference information from each respondent. Concerning this, we used SC, BWS1 and BWS2 tasks in the survey to understand how

[7] A preliminary pilot survey conducted at Shanghai Hongqiao Airport where the HSR-air intermodal service was available suggested low chance of intercepting transfer passengers, low willingness of outbound passengers to participate in the survey, and little knowledge about HSR-air intermodality of the participants. This also explains why we instead collected data at Pudong International Airport for the formal survey as it was much easier to approach transfer passengers there.

people react to the relatively new integrated HSR-air mode.

We collected data at Pudong International airport in Shanghai in January 2017. A total of 123 respondents answered 8 SC tasks, 7 BWS1 tasks and 8 BWS2 tasks. The SC component repeatedly asked participants to choose the most favourable alternative including the new HSR-air alternative. The BWS1 tasks examined the relative weight of all the 7 attributes involved in the SC tasks. The BWS2 tasks focused on the relative attractiveness of 14 attribute levels across 4 attributes of interest.

A detailed description of survey background, socio-demographic composition, SC experimental design, and descriptive analysis on the SC data can be found in Song, Hess, and Dekker (2018). All the respondents were shown tasks in the order of SC, BWS1 and BWS2, thus any ordering effects cannot be addressed in our study. We did so to ensure that respondents would be aware of the choice scenarios and the meaning of attributes involved in the SC tasks when they responded to the BWS1 and BWS2 tasks.

## 3.2. *SC tasks*

The context of the SC tasks is framed in the following way:

- a passenger is travelling from a domestic origin O to an overseas destination D;
- direct flights from O to D are unavailable;
- a passenger from O to D needs to travel via Shanghai;
- a passenger can only travel by air between Shanghai and D.

Four alternatives were shown to respondents, namely car-air, air-air, separated HSR-air and integrated HSR-air. As shown in Fig. 2, we denote the first leg between O and Shanghai as the "minor leg" on which various modes are available, and the second leg between Shanghai and D as the "major leg" where air is the only option. Car-air means using car on the minor leg and using flight on the major leg; air-air means taking a connecting flight; separated HSR-air refers to the traditional way of purchasing air and HSR tickets separately; integrated HSR-air refers to the new HSR-air intermodal service.

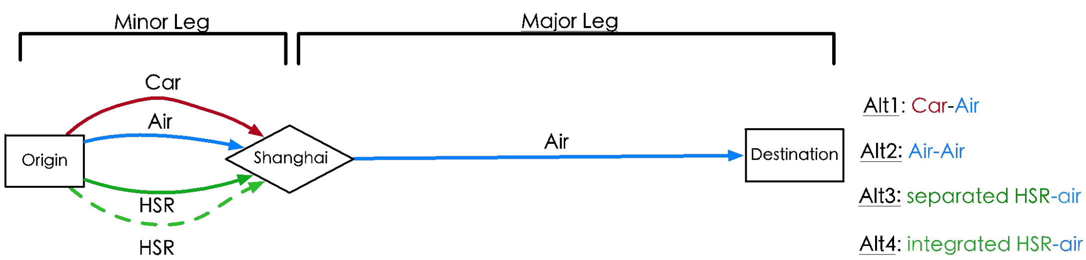


Figure 2.: Illustration of choice scenarios in the SC survey.

The SC survey was generated through a *D*-efficient design (Rose and Bliemer 2007) in Ngene (Metrics 2012). Each respondent was presented with 8 SC tasks in a randomised order, giving a total of 984 stated choice observations. Fig. 3 shows an example of the SC tasks. A total of 7 attributes were incorporated, including minor time, connection time, transfer time, delay protection, ticket integration, luggage integration and travel cost. Minor time gives the time spent on the minor leg; transfer time denotes the time spent on transferring between the minor leg and the major leg;[8] and connection time means the time spent on waiting and going through various procedures (e.g. security check-in, luggage check-in) at the departure airport of the major leg. Travel cost gives the total expenditure for the journey, and delay protection indicates to what extent a respondent would be compensated in case of delay on the minor leg. Ticket integration and luggage integration are two attributes describing the extent of integration of the ticketing systems and luggage-handling systems between the HSR side and the air side, of which the detailed levels can be found in Table 2.

| | Car-air | Air-air | Separated HSR-air | Integrated HSR-air |
|---|---|---|---|---|
| Travel cost | ¥1,250 | ¥1,050 | ¥1,150 | ¥1,250 |
| Minor time | 5h | 1.5h | 2.5h | 2.5h |
| Transfer time | *0h* | 0h | 1.5h | 1.5h |
| Connection time | *1.5h* | 4h | 1.5h | 2.5h |
| Delay protection | *None* | Free flight change | *None* | 50% discount on changing flight |
| Ticket integration | - | • *Book together*<br>• *Fixed-time flight on minor leg*<br>• *Easy collection* | • *Book separately*<br>• *Fixed-time train on minor leg*<br>• *No easy collection* | • Book together<br>• Fixed-time train on minor leg<br>• Easy collection |
| Security check and luggage integration | - | • Two security checks<br>• No integrated luggage handling system | • *Two security checks*<br>• *No integrated luggage handling system* | • One security check<br>• Integrated luggage handling system available |

Figure 3.: Example of SC tasks.

From the SC observations, we find that the integrated HSR-air alternative was most frequently chosen (41.57%), followed by the separated HSR-air alternative (26.42%), whereas car-air was selected for the least number of times (9.35%), which indicates relatively strong attractiveness of the integrated service and its potential market.

[8] Transfer time has three levels: it takes a value of 0min to indicate a seamless transfer in the same transport hub and takes the level of either 45min or 90min to suggest a transfer between two different hubs.

### 3.3. *BWS Case 1 tasks*

The BWS1 section required respondents to choose the attributes that they weighted the most and the least in each task. A balanced incomplete block design (BIBD) was adopted to generate the BWS1 experiment which could ensure each attribute occurred the same number of times and co-occurred with any other attribute the same number of times across all the choice tasks (Louviere, Flynn, and Marley 2015). In our survey, 7 attributes were assigned into 7 randomly-displayed BWS1 tasks, each with 4 attributes. Consequently, each attribute was shown to each respondent 4 times and each pair of attributes occurred twice. Fig. 4 shows an example of the BWS1 tasks.

| Most | If you are going to buy an integrated HSR-air service, what factors do you consider as the most important and least important? | Least |
|---|---|---|
| ☐ | Minor time | ☐ |
| ☐ | Delay protection | ☐ |
| ☐ | Connection time | ☐ |
| ☐ | Travel cost | ☐ |

Figure 4.: Example of BWS1 tasks.

An easy way to analyse BWS data is to compute the simple best-minus-worst (B-W) scores for each attribute.[9] Table 1 summarises the simple B-W score for each attribute averaged across respondents in a descending order as well as the standard deviation (s.d.) of individual-level simple B-W scores for each attribute. A higher B-W score means greater weight to the corresponding attribute in deciding whether to buy an integrated HSR-air option. These scores provide a straightforward implication that minor time and ticket integration mattered the least, whereas connection time and travel cost are the two attributes that mattered the most by the sample. The standard deviations of B-W scores suggest that respondents gave more diverse weight to the time-unrelated attributes than to time-related attributes. Minor time has the lowest B-W scores and is the attribute with the second lowest standard deviation of B-W scores, indicating that it was universally considered of limited importance. This is understandable as our survey was based in Shanghai and its nearby regions which could be reached by HSR or air from Shanghai within a relatively short period of time.

[9]Simple best-minus-worst scores can be obtained by subtracting the total count of an item being chosen as the worst from the total count the same item being chosen as the best across all BWS choice tasks and across all respondents (Louviere, Flynn, and Marley 2015). Since each attribute appeared 4 times per person in our case, the simple B-W score averaged at the individual-level is between -4 and 4.

Table 1.: Average simple B-W scores and standard deviation for BWS1 data

| Attribute | B-W score | s.d. | Score ranking |
|---|---|---|---|
| CT (connection time) | 0.37 | 2.00 | 1 |
| TC (travel cost) | 0.33 | 2.49 | 2 |
| DP (delay protection) | 0.29 | 2.35 | 3 |
| TT (transfer time) | 0.23 | 1.77 | 4 |
| LI (luggage integration) | 0.16 | 2.61 | 5 |
| TI (ticket integration) | -0.47 | 2.27 | 6 |
| MT (minor time) | -0.90 | 1.77 | 7 |

### 3.4. *BWS Case 2 tasks*

The BWS2 section consisted of 8 tasks, each comprising the attribute levels which constituted the profile of the integrated HSR-air alternative in each SC task. Our BWS2 survey focused on four attributes, i.e. connection time, delay protection, ticket integration and luggage integration, such that each BWS2 task required respondents to select the most appealing and the least appealing from 4 available attribute levels.[10] We did not include the full set of attributes in the BWS2 tasks as in the SC or BWS1 tasks for the sake of reducing cognitive burden and zooming in on those relatively new attributes of HSR-air. As the latent *attribute importance* is not used to show the influence of an attribute in comparison with other attributes, but to explain part of the inter-individual preference heterogeneity, not presenting levels for the remaining three attributes would not affect the distributions or the impact of the latent *attribute importance* across individuals for the four attributes involved in the BWS2 tasks.

Fig. 5 gives an example of the BWS2 tasks, where different levels across different attributes were evaluated on a common scale rather than being compared within an attribute, such that a respondent might prefer "having 50% off on a flight change" over "having an integrated luggage-handling system and one security check".

| Most | Given that the integrated HSR-air service costs 1250RMB, takes 2.5h on the minor (HSR) leg, and requires a transfer between Hongqiao HSR station and Pudong airport, which of the following are the most and the least appealing? | Least |
|---|---|---|
| ☐ | Connection time: 2.5h | ☐ |
| ☐ | 50% off on changing flight | ☐ |
| ☐ | Book together, fixed-time train on the minor leg and easy collection | ☐ |
| ☐ | Integrated luggage-handling and one security check | ☐ |

Figure 5.: Example of BWS2 tasks.

[10]The levels were always shown in the order of connection time, delay protection, ticket integration and luggage integration to reduce cognitive burden. Comparisons between levels within a same attribute were not allowed.

Table 2.: Summary of the attribute levels in BWS2 tasks

| # | Attribute level | Meaning | Numbers of respondents shown | Times available | Times as the best | Times as the worst |
|---|---|---|---|---|---|---|
| 1 | conn150 | Connection time is 2.5h | 123 | 235 | 32 | 53 |
| 2 | conn180 | Connection time is 3h | 111 | 172 | 15 | 83 |
| 3 | conn210 | Connection time in 3.5h | 123 | 280 | 25 | 97 |
| 4 | conn270 | Connection time is 4.5h | 74 | 162 | 2 | 93 |
| 5 | conn330 | Connection time is 5.5h | 87 | 135 | 1 | 103 |
| 6 | delay0 | No delay protection | 123 | 320 | 20 | 155 |
| 7 | delay1 | 50% off on changing flight should missing major-leg flight due to the delay on minor leg | 123 | 319 | 80 | 64 |
| 8 | delay2 | Changing flight for free should missing major-leg flight due to the delay on minor leg | 123 | 345 | 131 | 39 |
| 9 | tick1 | Booking tickets together, no easy collection, fixed-time train on the minor leg | 123 | 379 | 96 | 64 |
| 10 | tick2 | Booking tickets together, easy ticket collection available, fixed-time train on the minor leg | 123 | 324 | 76 | 56 |
| 11 | tick3 | Booking tickets together, eash ticket collection available, flexible train on the minor leg | 111 | 281 | 91 | 38 |
| 12 | lugg0 | No luggage integration, security checks required on both minor and major legs | 99 | 138 | 2 | 67 |
| 13 | lugg1 | Integrated luggage-handling system available, security checks required on both minor and major legs | 110 | 448 | 179 | 54 |
| 14 | lugg2 | Integrated luggage-handling system available, one security check required | 123 | 398 | 234 | 18 |

Overall, 14 different attribute levels were included in the BWS2 survey as listed in Table 2, including 5 levels of connection time, 3 levels of delay protection, 3 levels of ticket integration and 3 levels of luggage integration.

It should be noted that each item was not necessarily presented to all of the 123 respondents and did not occur with a same frequency. Thus, we calculate analytical B-W scores[11] to show relative attractiveness of the attribute levels among the sample. As shown in Table 3, we can see an increase in the analytical B-W scores as the level goes up for delay protection and luggage integration. However, for ticket integration, the scores are generally low and close to each other, indicating that the three levels of ticket integration were almost equally attractive to the respondents. One interesting thing is that connection time appears to be generally considered less attractive, regardless of which actual value it takes. This is understandable as connection time was considered as the most important factor in the BWS1 tasks, so that the respondents felt all the values of connection time presented in the BWS2 tasks to be unattractive.

The scores are used for descriptive analysis for better understanding the BWS1 and BWS2 data. All in all, we wish to study the correlation across the different datasets. The B-W scores themselves do not allow us to do so because we can only calculate the scores for BWS1 and BWS2 data independently, regardless of the calculation method we adopt. We need the joint model to simultaneously estimate on SC, BWS1 and BWS2 data and to explore the correlations among them.

Table 3.: Analytical B-W scores for BWS2 data at the sample level

| Attribute level | Analytical B-W score | Score ranking |
|---|---|---|
| conn150 | -0.18 | 8 |
| conn180 | -0.84 | 10 |
| conn210 | -0.53 | 9 |
| conn270 | -1.27 | 13 |
| conn330 | -1.97 | 14 |
| delay0 | -0.90 | 11 |
| delay1 | 0.10 | 7 |
| delay2 | 0.55 | 3 |
| tick1 | 0.17 | 5 |
| tick2 | 0.12 | 6 |
| tick3 | 0.38 | 4 |
| lugg0 | -1.02 | 12 |
| lugg1 | 0.57 | 2 |
| lugg2 | 1.22 | 1 |

[11] Analytical B-W scores can be obtained by $\ln\left(\frac{1+\frac{N_b-N_w}{N_x}}{1-\frac{N_b-N_w}{N_x}}\right)$, where $N_b - N_w$ is the simple B-W score and $N_x$ is the total times of the item being available, such that the score can rule out the impact of uneven occurrence of each attribute (Lipovetsky and Conklin 2014; Marley, Islam, and Hawkins 2016).

## 4. Case study: Model estimation

### 4.1. *Model specification*

The models in this paper were estimated in R using the flexible choice modelling package *Apollo* (Hess and Palma 2019), and 1000 MLHS draws (Hess, Train, and Polak 2006) were used in simulation. We used likelihood ratio tests to gradually improve the model specification and select the model offering the best fit while also taking into account the risk of over-fitting as well as behavioural interpretation of the modelling results. We also removed some insignificant variables due to small sample size and continuously checked the impact on willingness-to-pay estimates. This section describes the final specification of the joint ICLV model we have found with the best information criterion (i.e. Akaike Information Criterion and Bayesian Information Criterion), which can best balance between log-likelihood and behavioural insights while keeping the risk of over-fitting at a relatively low level..

#### *4.1.1. Structural equations*

After regressing the BWS1 individual-specific simple B-W scores of each attribute on different socio-demographic characteristics, the adopted structural equations for the 7 latent variables of attribute importance $\alpha_{nk}$ in Eq. 1 are defined as:[12]

$$\begin{aligned}
\alpha_{n,MT} &= \eta_{n,MT},\ (k = \text{Minor Time})\\
\alpha_{n,CT} &= \eta_{n,CT},\ (k = \text{Connection Time})\\
\alpha_{n,TT} &= \eta_{n,TT} + \omega_{TT,age>45} \cdot Z_{age>45},\ (k = \text{Transfer Time})\\
\alpha_{n,DP} &= \eta_{n,DP} + \omega_{DP,male} \cdot Z_{male},\ (k = \text{Delay Protection})\\
\alpha_{n,TI} &= \eta_{n,TI} + \omega_{TI,age>35} \cdot Z_{age>35},\ (k = \text{Ticket Integration})\\
\alpha_{n,LI} &= \eta_{n,LI} + \omega_{LI,age>45} \cdot Z_{age>45},\ (k = \text{Luggage Integration})\\
\alpha_{n,TC} &= \eta_{n,TC} + \omega_{TC,reimbursed} \cdot Z_{reimbursed},\ (k = \text{Travel Cost})
\end{aligned}\quad, \tag{13}$$

where $\eta_{nk}$ follows a standard Normal distribution among respondents. All socio-demographic variables used are rescaled to be centred on 0. We have not found suitable socio-demographics for the determinants of the latent attribute importance of minor time and connection time. Thus $\alpha_{n,MT}$ and $\alpha_{n,CT}$ are assumed to be purely random.

[12] For the sake of consistency, in section 4, parameters on attributes are notated with subscripts of the capital initials of the attributes as shown in Table 1, and parameters on attribute levels are represented with subscripts of the abbreviation of the attribute levels in lower case as listed in Table 3.

#### 4.1.2. Choice model for SC data

For normalisation purposes, the alternative-specific constant $\delta_i$ for the integrated HSR-air alternative is fixed to 0 while the other 3 alternative-specific constants are estimated. We assume $\tau_{MT} = 0$ to avoid over-specification since minor time acts as the base in the measurement model for BWS1 data and was not included in the BWS2 survey.

Minor time, connection time and travel cost are treated as continuous variables. The remaining four attributes are treated as categorical variables, with the lowest level of each being the base in dummy coding. The sensitivity coefficients for these attributes in the stated choice component in Eq. 3 are denoted in detail as:

$$\begin{aligned}
\beta_{n,MT} &= -e^{\mu_{ln(-\beta_{MT})}+\sigma_{ln(-\beta_{MT})}\xi_{n,MT}} \\
\beta_{n,CT} &= -e^{\tau_{CT}\alpha_{n,CT}} \cdot e^{\mu_{ln(-\beta_{CT})}+\sigma_{ln(-\beta_{CT})}\xi_{n,CT}} \\
\beta_{n,tran45\&90min} &= -e^{\tau_{TT}\alpha_{n,TT}} \cdot e^{\kappa_{TT,age>45}Z_{age>45}} \cdot e^{\mu_{ln(-\beta_{tran45\&90min})}+\sigma_{ln(-\beta_{TT})}\xi_{n,TT}} \\
\beta_{n,delay1\&2} &= \quad e^{\tau_{DP}\alpha_{n,DP}} \cdot e^{\kappa_{DP,male}Z_{male}} \cdot e^{\mu_{ln(\beta_{delay1\&2})}+\sigma_{ln(\beta_{DP})}\xi_{n,DP}} \\
\beta_{n,lugg1\&2} &= \quad e^{\tau_{LI}\alpha_{n,LI}} \cdot e^{\kappa_{LI,age>45}Z_{age>45}} \cdot e^{\mu_{ln(\beta_{lugg1\&2})}+\sigma_{ln(\beta_{LI})}\xi_{n,LI}} \\
\beta_{n,TC} &= -e^{\tau_{TC}\alpha_{n,TC}} \cdot e^{\kappa_{TC,reimbursed}Z_{reimbursed}} \cdot e^{\mu_{ln(-\beta_{TC})}+\sigma_{ln(-\beta_{TCx67})}\xi_{n,TC}}
\end{aligned}, \tag{14}$$

such that $\beta_{n,MT}$, $\beta_{n,CT}$ and $\beta_{n,TC}$ measure the marginal utilities, while $\beta_{n,tran45\&90min}$, $\beta_{n,delay1\&2}$, and $\beta_{n,lugg1\&2}$ give the relative utility against the corresponding base levels, which are *tran0min*, *delay0*, and *lugg0* in respective. The higher two levels for each are merged for estimation in our final specification as they are found not significantly different from each other. The final specification excludes the attribute of ticket integration from the utility function for the SC data, as it is found to contribute little to the utility functions. However, ticket integration is still used in the measurement models. Finally, parameters of $\kappa_{DP,male}$, $\kappa_{TC,reimbursed}$ and $\tau_{DP}$ are set to zero in the final specification as they were insignificant. Besides, although we have found suitable socio to explain transfer time (i.e. $Z_{age>45}$), the model with the indirect impact of $Z_{age>45}$ becomes insignificant once the direct impact is added. Hence, in the final specification, we drop the indirect impact by fixing $\omega_{TT,age>45} = 0$ and keep the direct impact of age on transfer time by estimating $\kappa_{TT,age>45}$.

#### 4.1.3. Measurement models for BWS1 data and BWS2 data

For the BWS1 data, all the 7 attributes shown in the SC survey are examined, i.e. minor time, connection time, transfer time, delay protection, ticket integration, luggage integration and travel cost. Minor time acts as the base, with relevant parameters $\delta_{MT|1}$ and $\zeta_{MT|1}$ normalised to 0. For the BWS2 data, connection time, delay protection, ticket integration and luggage integration are the four attributes of interest. Connection time is treated as a continuous variable and $x_{CT,nm|2}$ can take the value of 150min, 180min, 210min, 270min or 330min. The remaining three attributes are regarded as categorical variables, with level *delay0*, *tick1* and *lugg0* being the lowest (base) levels for delay protection, ticket integration and luggage integration in respective. The attribute level *delay0* is selected as the base in the measurement model for BWS2 data, with the baseline attractiveness $\phi_{delay0|2}$ fixed to 0 for normalisation.

### 4.2. *Estimation results*

For comparison, we estimated the corresponding reduced form mixed multinomial logit (MMNL) model for the SC data alone, i.e. setting $\tau = 0$, $\forall k$ (Vij and Walker 2016). The estimates of the MMNL model are shown alongside the estimates of the choice model component of the joint ICLV model in Table 4. In both models, the travel cost variable was scaled by 6.9, such that the value-of-time is expressed in the \$/min[13].

Since the ICLV model explains three different types of responses, the log-likelihood for the whole model in ICLV model ($LL(total) = -4445.339$) is much lower than the log-likelihood of the SC component alone. Meanwhile, the log-likelihood of the choice model component on the SC data of the ICLV model ($LL(SC) = -1060.453$) is slightly inferior to that of the MMNL model ($LL = -1057.396$), which is consistent with the discussions by Vij and Walker (2016). Indeed, the ICLV model needs to explain not only the SC data but also the extra BWS1 and BWS2 data, and it is then impossible for the ICLV model to outperform the reduced form MMNL model. Notwithstanding this, our joint ICLV model appears to provide more behavioural explanations than the reduced form MMNL model does. The $\tau$ estimates suggest significant roles of the latent variables of *attribute importance* in scaling sensitivities for all the non-cost attributes where applicable.

The MMNL model and the ICLV model show similar preference patterns towards attributes. As shown in the upper part of Table 4, the most negative $\delta_{ca}$ implies that the car-air alternative is the least preferred option, all else being equal, whereas the air-air alternative ($\delta_{aa}$) and the separated HSR-air alternative ($\delta_{sha}$) are both slightly less preferred compared to the base alternative, i.e. the integrated HSR-air mode. Since Lognormal distributions are used, the more negative the underlying mean parameter $\mu_{\ln|\beta_k|}$ is, the smaller in magnitude the median of marginal utility is, which translates into a lower sensitivity to that attribute in the SC tasks. As to the standard deviations $\sigma_{\ln|\beta_k|}$, both models detect statistically significant random heterogeneity in sensitivities to all of the attributes. Regarding the direct impacts of socio-demographics in the utility functions, we can see from both models that $\kappa_{TT,age>45}$ is significant at the 95% confidence interval, suggesting that older respondents are more sensitive to transfer

[13]USD/CNY$\approx$ 6.9 during the period of data collection.

Table 4.: Estimates for the reduced form MMNL model and the choice model component of the ICLV model

| | **MMNL** | | **ICLV** | |
|---|---|---|---|---|
| **Log likelihood** | LL: -1057.396 | | LL (total): -4445.399<br>LL (SC): -1060.453 | |
| | **est** | **t-rat(0)** | **est** | **t-rat(0)** |
| $\delta_{ca}$ | -3.210 | -7.49 | -3.081 | -6.91 |
| $\delta_{aa}$ | -0.411 | -1.73 | -0.439 | -2.04 |
| $\delta_{sha}$ | -0.622 | -3.30 | -0.738 | -3.60 |
| $\mu_{ln(-\beta_{MT})}$ | -5.243 | -16.51 | -5.441 | -14.26 |
| $\mu_{ln(-\beta_{CT})}$ | -4.527 | -37.69 | -4.596 | -38.62 |
| $\mu_{ln(-\beta_{tran45\&90min})}$ | -0.900 | -2.44 | -1.009 | -1.85 |
| $\mu_{ln(\beta_{delay1\&2})}$ | -1.342 | -2.29 | -2.157 | -2.42 |
| $\mu_{ln(\beta_{lugg1\&2})}$ | -0.729 | -2.32 | -1.096 | -2.10 |
| $\mu_{ln(-\beta_{TC})}$ | -4.181 | -22.02 | -4.265 | -14.51 |
| $\sigma_{ln(-\beta_{MT})}$ | -0.558 | -4.02 | -0.881 | -3.62 |
| $\sigma_{ln(-\beta_{CT})}$ | -0.517 | -6.11 | -0.409 | -5.02 |
| $\sigma_{ln(-\beta_{TT})}$ | 1.327 | 5.01 | 1.028 | 4.08 |
| $\sigma_{ln(\beta_{DP})}$ | -1.203 | -2.12 | -1.818 | -3.71 |
| $\sigma_{ln(\beta_{LI})}$ | -1.331 | -6.35 | -1.246 | -5.25 |
| $\sigma_{ln(-\beta_{TC})}$ | -0.622 | -3.75 | -0.486 | -2.81 |
| $\kappa_{TT,age>45}$ | 1.669 | 3.73 | 1.468 | 2.54 |
| $\kappa_{DP,male}$ | 0.000 | - | 0.000 | - |
| $\kappa_{LI,age>45}$ | 0.947 | 1.57 | 1.252 | 2.18 |
| $\kappa_{TC,reimbursed}$ | 0.000 | - | 0.000 | - |
| $\tau_{CT}$ | | | 0.233 | 2.37 |
| $\tau_{TT}$ | | | 0.335 | 2.59 |
| $\tau_{DP}$ | | | 0.000 | - |
| $\tau_{LI}$ | | | 0.701 | 4.49 |
| $\tau_{TC}$ | | | 0.334 | 1.21 |

time and dislike long transfer time more than young people do. Meanwhile, although $\kappa_{LI,age>45}$ in the MMNL model is only significant at the 80% confidence interval, we can still infer from $\kappa_{LI,age>45}$ in the ICLV model, which is significant at the 95% confidence interval, that older passengers can derive higher utility from better luggage integration than young people do.

In the left part of Table 5, the constant $\delta_{|1}$ represents the mean of the weight to the associated attribute among the sample in the BWS1 data. It could be noticed that, with minor time normalised to 0, connection time, delay protection and transfer time are positioned at the higher end of the underlying weighting scale, followed by travel cost and luggage integration. Regarding the scalars in the worst choice stage shown in the down left of Table 5, $\lambda_{CT|1}$ (t-rat(1)=-4.27) is the only one which is significantly different from 1, suggesting that scaling difference between the worst choice stage and the best choice stage only exists for the attribute of connection time. Since $\lambda_{CT|1}$ is much lower than 1, it suggests that the model has less noise in explaining the choices in the best choice stage than in the worst choice stage for the attribute of connection time.

Table 5.: Estimates of the measurement models for the BWS1 and BWS2 data using the MaxDiff models with scale difference

| **BWS1** | | | | **BWS2** | | | |
|---|---|---|---|---|---|---|---|
| | **est** | **t-rat(0)** | **t-rat(1)** | | **est** | **t-rat(0)** | **t-rat(1)** |
| $\delta_{MT\|1}$ | 0 (base) | - | - | $\delta_{CT\|2}$ | 4.151 | 4.06 | - |
| $\delta_{CT\|1}$ | 1.271 | 5.23 | - | $\gamma_{CT\|2}$ | -0.015 | -3.86 | - |
| $\delta_{TT\|1}$ | 0.920 | 4.22 | - | $\phi_{delay0\|2}$ | 0 (base) | - | - |
| $\delta_{DP\|1}$ | 1.071 | 3.21 | - | $\phi_{delay1\|2}$ | 2.008 | 5.54 | - |
| $\delta_{TI\|1}$ | 0.311 | 1.29 | - | $\phi_{delay2\|2}$ | 2.601 | 6.25 | - |
| $\delta_{LI\|1}$ | 0.738 | 2.37 | - | $\phi_{tick1\|2}$ | 1.956 | 4.86 | - |
| $\delta_{TC\|1}$ | 0.899 | 3.44 | - | $\phi_{tick2\|2}$ | 2.201 | 5.34 | - |
| | | | | $\phi_{tick3\|2}$ | 2.536 | 5.93 | - |
| | | | | $\phi_{lugg0\|2}$ | -0.102 | -0.33 | - |
| | | | | $\phi_{lugg1\|2}$ | 2.437 | 5.75 | - |
| | | | | $\phi_{lugg2\|2}$ | 3.432 | 7.60 | - |
| $\lambda_{MT\|1}$ | - | - | - | $\lambda_{MT\|2}$ | - | - | - |
| $\lambda_{CT\|1}$ | 0.255 | - | -4.27 | $\lambda_{CT\|2}$ | 0.992 | 4.11 | -0.03 |
| $\lambda_{TT\|1}$ | 0.600 | - | -1.17 | $\lambda_{TT\|2}$ | - | - | - |
| $\lambda_{DP\|1}$ | 0.751 | - | -0.98 | $\lambda_{DP\|2}$ | 0.815 | 7.18 | -1.63 |
| $\lambda_{TI\|1}$ | 1.171 | - | 0.48 | $\lambda_{TI\|2}$ | 0.691 | 5.41 | -2.42 |
| $\lambda_{LI\|1}$ | 1.018 | - | 0.06 | $\lambda_{LI\|2}$ | 0.755 | 6.59 | -2.13 |
| $\lambda_{TC\|1}$ | 1.411 | - | 0.95 | $\lambda_{TC\|2}$ | - | - | - |

The right part of Table 5 shows estimates for the baseline attractiveness of each attribute level in the BWS2 data. Focusing on $\phi_{|2}$, it can be inferred that compared to ticket integration, delay protection and luggage integration are associated with overall larger steps in attractiveness when moving from a poorer level to a better level, which implies that respondents might be indifferent to variations in ticket integration. This is in line with the discoveries in the SC data and the BWS1 data as well as the preliminary findings in the normalised B-W scores in the BWS2 data. As to the attribute-specific scalars shown in the down right of Table 5, only ticket integration $\lambda_{TI|2}$ (t-rat(1)=-2.42) and luggage integration $\lambda_{LI|2}$ (t-rat(1)=-2.13) are significantly different from 1. Being smaller than 1, $\lambda_{TI|2}$ and $\lambda_{LI|2}$ suggest stronger random error in the worst choice stage for these two attributes than in the best choice stage.

Now we turn to Table 6 to jointly examine all the impact factors of latent attribute importance in the choice model (i.e. $\tau$) as well as in the two MaxDiff-based measurement models (i.e. $\zeta_{|1}$ and $\zeta_{|2}$). The estimation results confirm our hypothesis. Except for $\tau_{TC}$, all the impact factors in the choice model and the measurement models are positive and significant where applicable. Thus, choices are made in a consistent way across different types of surveys. An increase in the latent variable would result in a stronger sensitivity to the associated attribute in the SC data, an increased probability that the attribute of interest is positioned to the higher end on the weighting scale in the BWS1 data, and a wider attractiveness gap between levels of the concerned attribute in the BWS2 data.

An exception arises for travel cost, where $\tau_{TC}$ is insignificant (est=0.334, t-rat(0)=1.21), whereas the same latent attribute importance plays a strong and significant role in BWS1 tasks (est=2.210, t-rat(0)=5.66). It is also worth noting that delay protection is related to cost as well, and that positive and significant impact of the corresponding latent attribute importance is found in both the BWS1 and BWS2 data, but not in the SC data, i.e. as mentioned earlier, $\tau_{DP}$ is fixed to 0 in this final specification as little influence from the latent attribute importance could be found on scaling the sensitivity to delay protection in the SC data. This implies a lack of consistency for the attributes related to cost between SC and BWS1/2 data, which is in accordance with and complements the findings in Balbontin, Ortúzar, and Swait (2015), where the sensitivity of an attribute related to cost, i.e. rent, was estimated to be inconsistent between the SC and BWS2 data. It might be due to the fact that choices in the SC experiment were made based on detailed choice contexts and level values of different attributes of each alternative in multi-alternative settings, while this information was not available in the BWS1 experiment where respondents' awareness and past experience of each attribute would influence their evaluation of the attributes (Louviere and Islam 2008; Mueller, Lockshin, and Louviere 2010). In this context, compared to the other non-cost attributes, it might be more difficult to assess the importance of the cost-relevant attributes and to trade off between cost and the other non-cost attributes without knowing the actual levels for all the available options in the choice set. Consequently, the role of the latent *attribute importance* is not significant in explaining the preference variations for cost-related attributes across individuals in the SC data, but is more prominent in the BWS1/2 data.

Combining the estimates $\omega$ in the structural equations and the impact factors for

Table 6.: Estimates in the structural equations and impact factors of latent attribute importance in the choice model and the BWS1/2 measurement models

| Structural equations | | | SC data | | | BWS1 data | | | BWS2 data | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | est | t-rat(0) | | est | t-rat(0) | | est | t-rat(0) | | est | t-rat(0) |
| $\omega_{MT}$ | - | - | $\tau_{MT}$ | - | - | $\zeta_{MT\|1}$ | 0 (base) | - | $\zeta_{MT\|2}$ | - | - |
| $\omega_{CT}$ | - | - | $\tau_{CT}$ | 0.233 | 2.37 | $\zeta_{CT\|1}$ | 0.659 | 2.03 | $\zeta_{CT\|2}$ | 0.373 | 9.37 |
| $\omega_{TT,age>45}$ | 0.000 | - | $\tau_{TT}$ | 0.335 | 2.59 | $\zeta_{TT\|1}$ | 1.211 | 4.50 | $\zeta_{TT\|2}$ | - | - |
| $\omega_{DP,male}$ | -0.863 | -2.71 | $\tau_{DP}$ | 0.000 | - | $\zeta_{DP\|1}$ | 2.067 | 3.40 | $\zeta_{DP\|2}$ | 0.519 | 3.25 |
| $\omega_{TI,age>35}$ | 0.868 | 3.97 | $\tau_{TI}$ | - | - | $\zeta_{TI\|1}$ | 1.683 | 4.34 | $\zeta_{TI\|2}$ | 0.371 | 3.94 |
| $\omega_{LI,age>45}$ | 1.191 | 2.66 | $\tau_{LI}$ | 0.701 | 4.49 | $\zeta_{LI\|1}$ | 2.160 | 5.29 | $\zeta_{LI\|2}$ | 0.530 | 4.80 |
| $\omega_{TC,reimbursed}$ | -0.625 | -3.36 | $\tau_{TC}$ | 0.334 | 1.21 | $\zeta_{TC\|1}$ | 2.210 | 5.66 | $\zeta_{TC\|2}$ | - | - |

latent attribute importance, the positive $\omega_{TI,age>35}$ and $\omega_{LI,age>45}$ and the negative $\omega_{TC,reimbuised}$ show that older people think ticket integration and luggage integration to be of greater importance than young people do, while passengers who get reimbursed perceive lower importance for travel cost than those who need to pay for the travel on their own. The negative and significant $\omega_{DP,male}$ suggests that male passengers find delay protection less important than female passengers do. Parameter $\omega_{TT,age>45}$ are fixed to 0 and not estimated in the final specification because of its very low significance. We can further look back into Table 4, where $\kappa_{TT,age>45}$ and $\kappa_{LI,age>45}$ are the only two statistically significant $\kappa$ parameters. We can therefore deduce that respondents' age mainly plays an independently direct role in scaling the marginal utility of transfer time, whereas age affects the marginal utility of luggage integration both directly and indirectly via the latent variable. The remaining socio-demographic characteristics involved in $\omega$ influence stated choice behaviour mainly through the latent variables of attribute importance.

Finally, we shed some light on willingness-to-pay (WTP) in the SC data with and without the additional information gained from the BWS1 and BWS2 data in Table 7. We first calculated the distributions of marginal utilities for all the attributes, taking into account of the roles of latent attribute importance and socio-demographic characteristics in the ICLV model and the role of socio-demographic characteristics in the reduced form MMNL model, i.e. marginal utilities $\beta_{nk}$ are given by $e^{\tau_k \alpha_{nk}} e^{\kappa_k Z_n} \beta^\star_{nk}$ in the ICLV model and by $e^{\kappa_k Z_n} \beta^\star_{nk}$ in the MMNL model, where $\beta^\star_{nk} = e^{\mu_{\ln\beta_k} + \sigma_{\ln\beta_k} \cdot \xi_{nk}}$. We then calculated the ratio against the marginal utility of travel cost for each of the remaining attributes for each draw, which is taken from the distributions of marginal utilities used in the estimation procedure, enabling us to obtain the WTP distributions for all the attributes except for travel cost through simulation (Hensher and Greene 2003; Sillano and de Dios Ortúzar 2005; Daly, Hess, and Train 2012).

We see some differences between the two models here, where we would argue that the ICLV findings are more realistic especially for transfer time. Indeed, in the ICLV model, going from a transfer time of 45 or 90 minutes to a seamless transfer has the same benefit as a reduction in connection time by 81.6 minutes at the mean. In the MMNL model, this would be 122.58 minutes, which seems unrealistic if we assume that transfer time should at best be as important as connection time. In addition, the standard deviations of the three categorical attributes, i.e. transfer time, delay protection, and luggage integration are relatively large in both models. This can be mainly attributed to the long tails of the Lognormal distributed WTP distributions as the marginal utilities for all the attributes follow Lognormal distributions. Hence, apart from regular statistics of mean and standard deviation, we also show the median and interquartile range of each WTP distribution. We can see an overall reduction in the median values, and a decrease in the interquartile range for all the attributes except for minor time when we move from the MMNL model to the ICLV model. This means that the spread of the distribution is smaller and and the values are more squeezed to the median for the ICLV model.

Table 7.: WTP estimates of the joint ICLV model and the reduced form MMNL model.

| models | attributes | sensitivities $\beta$ | | mean and percentiles of WTP distribution | | | | WTP changes against MMNL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mean | s.d. | mean | s.d. | median | interquartile range | mean | s.d. | median | interquartile range |
| ICLV | Minor Time | -0.006 | 0.007 | 0.54 | 0.78 | 0.31 | 0.48 | 10% | 59% | -11% | 17% |
| | Connection Time | -0.011 | 0.006 | 0.96 | 0.85 | 0.72 | 0.77 | -2% | -9% | 1% | -5% |
| | Transfer Time_45&90min | -0.738 | 1.429 | 62.72 | 146.51 | 25.47 | 50.34 | -32% | -55% | -2% | -22% |
| | Delay Protection_lv1&2 | 0.606 | 2.981 | 52.62 | 359.14 | 8.18 | 27.75 | 23% | 252% | -52% | -23% |
| | Luggage Integration_lv1&2 | 1.231 | 5.119 | 104.63 | 509.18 | 23.01 | 62.19 | 8% | 78% | -27% | -17% |
| | Travel Cost | -0.017 | 0.011 | - | - | - | - | - | - | - | - |
| MMNL | Minor Time | -0.006 | 0.004 | 0.49 | 0.49 | 0.35 | 0.41 | - | - | - | - |
| | Connection Time | -0.012 | 0.007 | 0.98 | 0.93 | 0.71 | 0.81 | - | - | - | - |
| | Transfer Time_45&90min | -1.160 | 3.581 | 91.80 | 328.10 | 26.08 | 64.19 | - | - | - | - |
| | Delay Protection_lv1&2 | 0.539 | 0.975 | 42.87 | 101.98 | 16.99 | 35.81 | - | - | - | - |
| | Luggage Integration_lv1&2 | 1.221 | 2.833 | 97.05 | 285.32 | 31.44 | 75.02 | - | - | - | - |
| | Travel Cost | -0.019 | 0.013 | - | - | - | - | - | - | - | - |

## 5. Conclusions

This research has looked at potential travel behaviour in the context of the introduction of a new travel mode, i.e. HSR-air intermodality. The need for better understanding the role of attributes (especially the new ones) in the new context entails collecting more behavioural information from each individual. Compared with adopting a longer SC survey, synthesising data from multiple types of preference elicitation approaches can reduce boredom caused by additional SC tasks and provide more robust explanation of the role that attributes play. The growing interest in BWS data has presented the potential of such data synthesis. Specifically, SC data allows us to analyse how respondents trade off between attributes and forecast demand, whereas BWS1 and BWS2 data helps in providing more behavioural insights about the role that attributes play. It needs to be noted that it is not the objective of this research to conclude which type of preference elicitation method is more correct.

Informed by the work of Hess and Hensher (2013), we adopt the notion of *attribute importance* and treat it as a latent variable, which acts as the connection amongst all the three types of data. The attribute-specific latent variable scales the marginal utility of the associated attribute in the choice model for the SC data. Meanwhile, it explains the weight of the attribute and scale the marginal attractiveness of attribute levels in the measurement models for the BWS1 data and the BWS2 data respectively.

This research has for the first time collected SC data together with more than one type of BWS data from the same respondents. Our work can provide researchers with practical guidance on applying BWS1 and (or) BW2 approaches in travel behaviour contexts, and insights of choice behaviour in different types of surveys. By simultaneously estimating on the SC, BWS1 and BWS2 data through the latent constructs of *attribute importance* in the ICLV model, we are able to examine the correlations of choice behaviour among these three different types of tasks at the individual level, which was not addressed in Balbontin, Ortúzar, and Swait (2015), without inducing the risk of endogeneity bias or measurement error which arose in Beck, Rose, and Greaves (2017). The use of BWS1 and BWS2 data in the measurement models of the ICLV model also provides richer behavioural information than the earlier work by Hess and Hensher (2013), where stated attribute attendance and attribute rankings were used.

Overall, our joint model shows that attribute importance can link the SC, BWS1 and BWS2 data, indicating the benefit of improving behavioural explanation by combining the BWS data with SC data. We found a high level of consistency with respect to the impact of the underlying perceived *attribute importance* on decision-making in different tasks is significantly demonstrated. The estimation results imply that an increase in attribute importance results in a stronger sensitivity to that attribute in the SC tasks, more overall weight to that attribute in the BWS1 tasks, and also wider attractiveness gaps between levels for that attribute in the BWS2 tasks. This is particularly true for non-cost attributes, including connection time, transfer time and luggage integration in our case. We have not found similar consistency for cost-relevant attributes, i.e. delay protection and travel cost, as the corresponding latent variables only impose significant impacts in the BWS1/2 data but not in the SC data. That is, we have not discovered a one-to-one relationship between different survey methods. As such, there remain some differences in how attribute importance is evaluated between SC, BWS1 and BWS2 data. We therefore think treating different survey methods as equivalent and interchangeable - for example using BWS1 method to determine which attributes to include in SC survey - can be risky.

The lack of one-to-one consistency between different types of data is understandable as SC tasks were conducted in multi-alternative settings. Meanwhile, the detailed information of attribute levels and (or) the information of other competing alternatives were not available in BWS1 tasks, and the competing alternatives were also not shown to respondents. Thus respondents would be more capable to make trade-offs among attributes based on the presented information in SC tasks, whereas their perceived importance of a given attributes in a BWS1/2 survey is more affected by personal experience etc. (Louviere and Islam 2008; Mueller, Lockshin, and Louviere 2010).

The finding that there is not a one-to-one relationship between the different types of data can also be due to the fact that selecting the best is different from selecting the worst, i.e. best choices are made under positive frames whereas worst choices are made within negative frames (Rose 2014; Giergiczny et al. 2017). Given these results, we suggest that researchers should not see BWS data as a replacement for SC data in preference elicitation research. It is of course feasible to use BWS tasks alongside SC tasks for better explanation of choices made in SC tasks, and this may be especially beneficial if the number of respondents is low. We acknowledge that Hawkins, Islam, and Marley (2018) suggested that the conclusion of best choices and worst choice being made in different ways in many studies were due to the inadequate data. They argued that respondents made best choices and worst choices in a same way (i.e. same utility parameters), while worst choices were usually associated with greater variance in the error term (i.e. scale heterogeneity existed between best choice stage and worst choice stage). In our paper, the best choice stage and worst choice stage share the same specification but with attribute-specific scale parameters imposed on the worst stage. This means that our model is more generic and flexible, enabling us to detect whether and which attribute has different scales between best and worst stages. The results suggested that only a subset of attributes influence decision-making differently on the worst stage in comparison to the best stage. Besides, we were using only a small sample of data, which in turn makes it difficult to adopt more complex model specification or to validate the conclusion raised by Hawkins, Islam, and Marley (2018). Regarding this, it is necessary and beneficial to replicate different methods in more research contexts.

The present work also has some limitations. Firstly, systematic order effects were not accounted for in our case study as respondents were all presented with choice tasks in the order of SC, BWS1 and BWS2. Secondly, due to the restriction of sample size, all the preference variations in the BWS1 and BWS2 tasks were attributed to latent attribute importance, and we did not incorporate random heterogeneity irrelevant to latent variables in out final specification. It would be worth applying our method on other larger joint datasets with more complicated specification of random heterogeneity, while at the same time achieving a balance with higher computational burden. Furthermore, we could test the non-linearity in sensitivity parameters on the utility functions for alternatives in the SC data.

## Acknowledgements

Fangqing Song acknowledges the support of the China Scholarship Council while Stephane Hess was supported by the European Research Council through the consolidator grant 615596-DECISIONS. We are grateful for the constructive comments received from the two anonymous reviewers who have helped us significantly improve this work.

## Appendix A. The descriptors of the notations used in section 2.

| | |
|---|---|
| $\alpha$ | Matrix, giving the latent *attribute importance* of each attribute perceived by each respondent. |
| $\alpha_n$ | Vector, giving the latent *attribute importance* of each attribute perceived by respondent $n$. |
| $\alpha_{nk}$ | Scalar, giving the latent *attribute importance* of attribute $k$ perceived by respondent $n$. |
| $\beta$ | Matrix, describing the marginal utility of each attribute for each respondent. |
| $\beta_n$ | Vector, describing the marginal utility of each attribute for respondent $n$. |
| $\beta_k$ | Vector, describing the marginal utility of attribute $k$ for each respondent. |
| $\beta_{nk}$ | Scalar, describing the marginal utility of attribute $k$ perceived by respondent $n$. |
| $(b,w)_{\|1}$ | Matrix, giving the choice (i.e. pair of the best attribute $b$ and the worst attribute $w$) for each respondent in each BWS1 choice task. |
| $(b,w)_{\|2}$ | Matrix, giving the choice (i.e. pair of the best attribute level $b$ and the worst attribute level $w$) for each respondent in each BWS2 choice task. |
| $B_{qnm\|c}$ | Scalar, denoting the "utility" (i.e. weight of an attribute or attractiveness of an attribute level) of item $q$ in the "best" stage for respondent $n$ as shown in BWS task $m$ and BWS type $c$ (i.e. $c = 1$ stands for BWS1 and $c = 2$ stands for BWS2). |
| $BW_{(q,j)nm\|c}$ | Scalar, denoting the "utility" difference between item $q$ and item $j$ for respondent $n$ as shown in BWS case $c$ task $m$, with $q$ standing for the best and $j$ standing for the worst in the pair $(q, j)$. |
| $\delta_i$ | Scalar, a constant in the utility function for alternative $i$ in SC tasks, which is generic across respondents and tasks. |
| $\delta_{k\|1}$ | Scalar, capturing the mean weight of attribute $k$ in BWS1 tasks, which is generic across respondents and tasks. |
| $\delta_{k\|2}$ | Scalar, a constant associated with attribute $k$ in BWS2 tasks (only apply to the situation where $k$ is a continuous variable). |
| $\eta_{nk}$ | Describing the standard Normal error term for respondent $n$ and attribute $k$. |
| $\gamma_{k\|2}$ | Scalar, capturing the baseline marginal attractiveness of the attribute levels of attribute $k$ (only apply to the situation where $k$ is a continuous variable). |

| | |
|---|---|
| $\kappa$ | Matrix, describing the impact of each socioeconomic variable on each attribute's marginal utility. |
| $\kappa_k$ | Vector, describing the impact of each socio-demographic variable on the marginal utility of attribute $k$. |
| $\lambda_{j\|c}$ | Scalar, capturing the scale difference between the "best" and the "worst" stage for item $j$ in BWS case $c$ tasks. |
| $L_k$ | Scalar, giving the total number of possible values that attribute $k$ can take in a BWS2 survey. |
| $\mu_{\ln \beta_k}$ | Scalar, capturing the mean of the underlying Normal distribution for $\beta_k$. |
| $\nu_{qjnm\|c}$ | Describing a standard extreme value type I error term operating at the level of the attribute (level) pair of $(q, j)$ for respondent $n$ in BWS case $c$ task $m$. |
| $\omega$ | Matrix, describing the impact of each socio-demographic variable on each attribute's corresponding latent *attribute importance*. |
| $\omega_k$ | Vector, measuring the impact of each socio-demographic variable on the latent *attribute importance* for attribute $k$. |
| $\phi_{k_l\|2}$ | Scalar, denoting the baseline attractiveness of level $l$ for attribute $k$ in BWS2 tasks (only apply to the situation where $k$ is a categorical 1 variable). |
| $\sigma_{\ln \beta_k}$ | Scalar, capturing the standard deviation of the underlying Normal distribution for $\beta_k$. |
| $\tau$ | Vector, describing the impact of each latent *attribute importance* on the corresponding attribute's marginal utility in the SC component. |
| $\tau_k$ | Scalar, describing how the marginal utility of attribute $k$ is affected by the corresponding *attribute importance* in the SC component. |
| $U_{int}$ | Scalar, representing the utility of alternative $i$ derived by respondent $n$ in SC task $t$. |
| $V_{int}$ | Scalar, representing the systematic utility of alternative $i$ for respondent $n$ in SC task $t$. |
| $\varepsilon_{int}$ | Describing the unobserved type I extreme value error of $U_{int}$. |
| $x_{int}$ | Vector, explanatory variables representing the $K$ attributes of alternative $i$ as shown to respondent $n$ in SC task $t$. |
| $x_{intk}$ | Scalar, the explanatory variable representing attribute $k$ of alternative $i$ as shown to respondent $n$ in SC task $t$. |
| $x_{knm\|2}$ | Scalar, denoting the level value that attribute $k$ takes for respondent $n$ in BWS2 task $m$. |
| $\xi_{nk}$ | Describing the value of a standard Normal distribution across respondents for attribute $k$ taken by respondent $n$. |

| | |
|---|---|
| $W_{jnm\|c}$ | Scalar, denoting the "utility" (i.e. weight of an attribute or attractiveness of an attribute level) of item $j$ in the "worst" stage for respondent $n$ as shown in BWS type $c$ task $m$. |
| $y$ | Matrix, giving the choice for each respondent in each stated choice task. |
| $y_{nt}$ | Scalar, giving the choice by respondent $n$ in stated choice task $t$. |
| $\zeta_{\|1}$ | Vector, describing the impact of each latent *attribute importance* on the corresponding attribute's weight in the BWS1 component. |
| $\zeta_{k\|1}$ | Scalar, describing how the weight of attribute $k$ is affected by the corresponding latent *attribute importance* in the BWS1 component. |
| 1 $\zeta_{\|2}$ | Vector, describing the impact of each latent *attribute importance* on the corresponding attribute levels' attractiveness in the BWS2 component. |
| $\zeta_{k\|2}$ | Scalar, describing how the level spacing for attribute $k$ in terms of attractiveness is affected by the corresponding latent *attribute importance* in the BWS2 component. |
| $Z$ | Matrix, giving the value of each socio-demographic variable for each respondent. |
| $Z_n$ | Vector, giving the value of each socio-demographic variable for respondent $n$. |

## References


Auger, Pat, Timothy M Devinney, and Jordan J Louviere. 2007. "Using best–worst scaling methodology to investigate consumer ethical beliefs across countries." *Journal of Business Ethics* 70 (3): 299–326.

Balbontin, Camila, J de D Ortúzar, and JD Swait. 2015. "A joint best–worst scaling and stated choice model considering observed and unobserved heterogeneity: An application to residential location choice." *Journal of choice modelling* 16: 1–14.

Beck, Matthew J, and John M Rose. 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.

Beck, Matthew J, John M Rose, and Stephen P Greaves. 2017. "I cant believe your attitude: a joint estimation of best worst attitudes and electric vehicle choice." *Transportation* 44 (4): 753–772.

Ben-Akiva, Moshe, Joan Walker, Adriana T Bernardino, Dinesh A Gopinath, Taka Morikawa, and Amalia Polydoropoulou. 2002. "Integration of choice and latent variable models." *Perpetual motion: Travel behaviour research opportunities and application challenges* 431–470.

Campbell, Danny, David A Hensher, and Riccardo Scarpa. 2011. "Non-attendance to attributes in environmental choice analysis: a latent class specification." *Journal of environmental planning and management* 54 (8): 1061–1076.

Carlsson, Fredrik. 2003. "The demand for intercity public transport: the case of business passengers." *Applied Economics* 35 (1): 41–50.

Coast, Joanna, Terry N Flynn, Chris Salisbury, Jordan Louviere, and Tim J Peters. 2006. "Maximising responses to discrete choice experiments." *Applied health economics and health policy* 5 (4): 249–260.

Collins, Andrew T, Michiel CJ Bliemer, and John M Rose. 2014. "Constrained stated choice experimental designs." .

Daly, Andrew, Stephane Hess, and Kenneth Train. 2012. "Assuring finite moments for willingness to pay in random coefficient models." *Transportation* 39 (1): 19–31.

Dumont, Jeffrey, Marek Giergiczny, and Stephane Hess. 2015. "Individual level models vs. sample level models: contrasts and mutual benefits." *Transportmetrica A: Transport Science* 11 (6): 465–483.

Dyachenko, Tatiana, Rebecca Walker Reczek, and Greg M Allenby. 2014. "Models of sequential evaluation in best-worst choice tasks." *Marketing Science* 33 (6): 828–848.

Finn, Adam, and Jordan J Louviere. 1992. "Determining the appropriate response to evidence of public concern: the case of food safety." *Journal of Public Policy & Marketing* 12–25.

Giergiczny, Marek, Thijs Dekker, Stephane Hess, and P Chintakayala. 2017. "Testing the stability of utility parameters in repeated best, repeated best-worst and one-off best-worst studies." *European Journal of Transport and Infrastructure Research* 17 (4): 457–476.

Hawkins, Guy E, Towhidul Islam, and AAJ Marley. 2018. "Like it or not, you are using one value representation." *Decision* .

Hensher, David A. 2006. "How do respondents process stated choice experiments? Attribute consideration under varying information load." *Journal of applied econometrics* 21 (6): 861–

878.

Hensher, David A, and William H Greene. 2003. "The mixed logit model: the state of practice." *Transportation* 30 (2): 133–176.

Hensher, David A, Corinne Mulley, and John M Rose. 2015. "Understanding the relationship between voting preferences for public transport and perceptions and preferences for bus rapid transit versus light rail." *Journal of Transport Economics and Policy (JTEP)* 49 (2): 236–260.

Hensher, David A, John Rose, and William H Greene. 2005. "The implications on willingness to pay of respondents ignoring specific attributes." *Transportation* 32 (3): 203–222.

Hensher, David A, and John M Rose. 2009. "Simplifying choice through attribute preservation or non-attendance: implications for willingness to pay." *Transportation Research Part E: Logistics and Transportation Review* 45 (4): 583–590.

Hess, Stephane, and David A Hensher. 2010. "Using conditioning on observed choices to retrieve individual-specific attribute processing strategies." *Transportation Research Part B: Methodological* 44 (6): 781–790.

Hess, Stephane, and David A Hensher. 2013. "Making use of respondent reported processing information to understand attribute importance: a latent variable scaling approach." *Transportation* 40 (2): 397–412.

Hess, Stephane, and David Palma. 2019. "Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application." *Journal of choice modelling* 32: 100170.

Hess, Stephane, Amanda Stathopoulos, Danny Campbell, Vikki ONeill, and Sebastian Caussade. 2013. "Its not that I dont care, I just dont care very much: confounding between attribute non-attendance and taste heterogeneity." *Transportation* 40 (3): 583–607.

Hess, Stephane, Kenneth E Train, and John W Polak. 2006. "On the use of a modified latin hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice." *Transportation Research Part B: Methodological* 40 (2): 147–163.

Lipovetsky, Stan, and Michael Conklin. 2014. "Best-worst scaling in analytical closed-form solution." *Journal of choice modelling* 10: 60–68.

Louviere, Jordan J, Terry N Flynn, and Anthony AJ Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Louviere, Jordan J, and Towhidul Islam. 2008. "A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best–worst scaling." *Journal of Business Research* 61 (9): 903–911.

Marley, AAJ, T Islam, and GE Hawkins. 2016. "A formal and empirical comparison of two score measures for best–worst scaling." *Journal of choice modelling* 21: 15–24.

Marley, Anthony AJ, and Jordan J Louviere. 2005. "Some probabilistic models of best, worst, and best–worst choices." *Journal of Mathematical Psychology* 49 (6): 464–480.

Marley, Anthony Alfred John, Terry N Flynn, and Jordan Joseph Louviere. 2008. "Probabilistic models of set-dependent and attribute-level best–worst choice." *Journal of Mathematical Psychology* 52 (5): 281–296.

Marti, Joachim. 2012. "A best–worst scaling survey of adolescents' level of concern for health and non-health consequences of smoking." *Social Science & Medicine* 75 (1): 87–97.

McFadden, Daniel. 2001. "Economic choices." *American economic review* 91 (3): 351–378.
McFadden, Daniel, et al. 1973. "Conditional logit analysis of qualitative choice behavior." .
Metrics, Choice. 2012. "Ngene 1.1. 1 User Manual & Reference Guide." *Sydney, Australia: ChoiceMetrics* .
Mueller, Simone, Larry Lockshin, and Jordan J Louviere. 2010. "What you see may not be what you get: Asking consumers what matters may not reflect what they choose." *Marketing Letters* 21 (4): 335–350.
Petrolia, Daniel R, Matthew G Interis, and Joonghyun Hwang. 2018. "Single-Choice, Repeated-Choice, and Best-Worst Scaling Elicitation Formats: Do Results Differ and by How Much?" *Environmental and Resource Economics* 69 (2): 365–393.
Potoglou, Dimitris, Peter Burge, Terry Flynn, Ann Netten, Juliette Malley, Julien Forder, and John E Brazier. 2011. "Best–worst scaling vs. discrete choice experiments: an empirical comparison using social care data." *Social science & medicine* 72 (10): 1717–1727.
Pullman, Madeleine E, Kimberly J Dodson, and William L Moore. 1999. "A comparison of conjoint methods when there are many attributes." *Marketing Letters* 10 (2): 125–138.
Rose, John M. 2014. *Interpreting discrete choice models based on best-worst data: a matter of framing*. Technical Report.
Rose, John M, and Michiel CJ Bliemer. 2007. "Stated preference experimental design strategies." In *Handbook of Transport Modelling: 2nd Edition*, 151–180. Emerald Group Publishing Limited.
Sillano, Mauricio, and Juan de Dios Ortúzar. 2005. "Willingness-to-pay estimation with mixed logit models: some new evidence." *Environment and Planning A* 37 (3): 525–550.
Song, Fangqing, Stephane Hess, and Thijs Dekker. 2018. "Accounting for the impact of variety-seeking: Theory and application to HSR-air intermodality in China." *Journal of Air Transport Management* 69: 99–111.
Train, Kenneth E. 2009. *Discrete choice methods with simulation*. Cambridge university press.
Vij, Akshay, and Joan L Walker. 2016. "How, when and why integrated choice and latent variable models are latently useful." *Transportation Research Part B: Methodological* 90: 192–217.