

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

3 Fangqing Song^{a,b}, Stephane Hess^{a,b} and Thijs Dekker^{a,b}

4 ^aInstitute for Transport Studies, University of Leeds, UK;

5 ^bChoice Modelling Centre, University of Leeds, UK

6 **ARTICLE HISTORY**

7 Compiled May 22, 2020

8 **ABSTRACT**

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

32 **KEYWORDS**

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

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

¹In 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.

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

3 This research is conducted in the context where a new travel mode, i.e. high-speed
4 rail(HSR)-air intermodality, is introduced. Since our interest is in predicting choices
5 (i.e. first preferences only), we adopt a traditional SC survey as it allows us to analyse
6 how respondents make trade-offs between attributes and forecast travel demand within
7 multi-alternative settings. A BWS Case 3 survey is not adopted for this purpose as it
8 combines both the best and the worst where existing studies show diverging views on
9 how consistent people are in choosing the best and the worst. Some found differences in
10 both utility parameters and scales between the two stages (Rose 2014; Giergiczny et al.
11 2017), notwithstanding contrary findings in Hawkins, Islam, and Marley (2018) that
12 suggested that the same utility parameters drive individuals' best and worst choices
13 despite a scale difference between best choices and worst choices. In addition to the
14 SC survey, BWS Case 1 and BWS Case 2 surveys are used as these two methods can
15 reflect how individuals are influenced by different attributes in relatively more direct
16 manners in single-alternative settings. As such, BWS Case 1 and BWS Case 2 data
17 serves as additional behavioural information to help in better explaining the role of
18 specific attributes in these choice decisions.²

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

30 The majority of studies comparing SC data and BWS Case 1 and (or) Case 2 data
31 have been conducted at the sample level (e.g. Louviere and Islam 2008; Potoglou
32 et al. 2011). Only Balbontin, Ortúzar, and Swait (2015) and Beck, Rose, and Greaves
33 (2017) have jointly analysed SC and BWS Case 2 data. However, there are some
34 remaining limitations associated with these two joint estimation studies. The former
35 lacks flexibility in model specifications as it assumes the impact of an attribute level
36 in the SC tasks to be equal (or a function of) to the impact of the same attribute
37 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.

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

4 In this paper, we put forward a flexible approach to jointly estimate SC, BWS Case
5 1 and BWS Case 2 data at the individual level while overcoming the shortcomings in
6 the literature. This approach is based on the assumption that responses to BWS Case
7 1, BWS Case 2 and SC tasks are all driven by a common underlying factor of perceived
8 *attribute importance*. We develop an Integrated Choice and Latent Variable (ICLV)
9 model (Ben-Akiva et al. 2002) where each attribute is associated with a specific latent
10 variable of *attribute importance*. The notion of *attribute importance* has previously
11 been put forward to challenge the decision heuristic of attribute non-attendance (Hen-
12 sher, Rose, and Greene 2005; Hensher 2006; Hensher and Rose 2009), arguing that
13 some people actually perceive reduced importance for an attribute in making stated
14 choices rather than completely ignoring it even if the respondents stated that they
15 did not take the associated attribute into account (Hess and Hensher 2010; Camp-
16 bell, Hensher, and Scarpa 2011; Hess et al. 2013). Our work adopts a similar strategy
17 as Hess and Hensher (2013), who use latent *attribute importance* to simultaneously
18 explain the responses to SC tasks and the responses to selected indicators, including
19 binary stated attribute attendance and stated attribute rankings. In our proposed
20 model, the indicators are replaced by BWS Case 1 and Case 2 data.

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

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

³BWS 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.

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

2 **2. Methodology**

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

6 **2.1. Model framework**

7 As mentioned in the Introduction, our model is developed based on the assumption
8 of correlation between SC responses and BWS1/2 responses. Latent variables are in-
9 troduced to capture the correlation and to simultaneously explain different types of
10 responses within a single ICLV framework. We follow the adoption of the notion *at-*
11 *tribute importance* from Hess and Hensher (2013) to represent latent variables for each
12 attribute as SC, BWS1 and BWS2 surveys all reveal people’s preferences towards var-
13 ious attributes in the decision-making process.⁴

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

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

⁴Please 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.

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

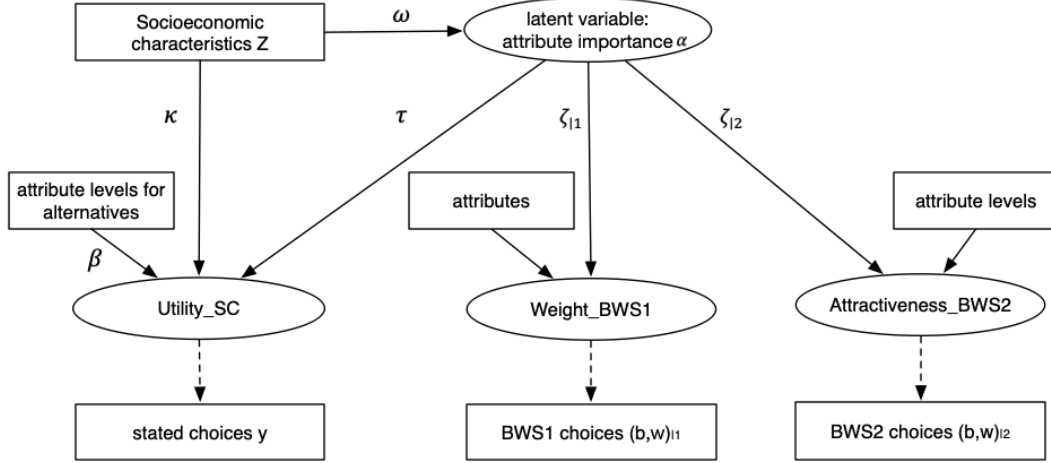


Figure 1.: Framework of the joint model.

6 2.2. Structural equations for latent variables

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

$$\alpha_{nk} = \omega'_k Z_n + \eta_{nk}, \quad (k = (1, \dots, K)), \quad (1)$$

11 where η_{nk} is a standard Normal error term and where the estimated vector of param-
 12 eters ω_k measures the impact of the socio-demographic characteristics on the latent
 13 variable. Note that Z_n is centred on 0, such that the latent variable α_{nk} has a mean
 14 of 0.

15 2.3. Stated choice model component

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

20 Let U_{int} in Eq. 2 represent the utility of alternative i for respondent n in stated

1 choice task t . U_{int} consists of a deterministic portion V_{int} (i.e. systematic utility), and
 2 an unobserved error term ε_{int} which is independently and identically distributed (IID)
 3 extreme value type I.

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

4 The term δ_i is an estimated alternative-specific constant (ASC) while $x_{int} =$
 5 $(x_{int1}, \dots, x_{intK})'$ is a vector of explanatory variables representing the K attributes
 6 of alternative i as shown to respondent n in SC task t , where the estimated vector
 7 $\beta_n = (\beta_{n1}, \dots, \beta_{nK})'$ captures the marginal utilities of these attributes. Hence, it is
 8 assumed that each attribute contributes to the utility of an alternative in an addi-
 9 tive manner, and that the marginal utility for each attribute is kept generic across
 10 alternatives.

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

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

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

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

17 Latent *attribute importance* is accommodated in an exponential form to act as
 18 a positive scalar on marginal utility where τ_k captures the degree of scaling (Hess
 19 and Hensher 2013). To avoid overstating the role of latent *attribute importance* in
 20 explaining heterogeneity in the SC data (Vij and Walker 2016), we let the socio-
 21 demographics Z_n which explain the latent variable α_{nk} in the structural equations also
 22 directly enter the marginal utility, where the vector κ_k measures the direct impacts
 23 from socio-demographics Z_n on the scaling of marginal utility. Additional random
 24 heterogeneity that is not linked to the latent variable is accommodated by specifying
 25 the underlying parameter, net of the influence of socio-demographics and the latent
 26 variable, to follow a Lognormal distribution. We then have that $\mu_{\ln \beta_k}$ and $\sigma_{\ln \beta_k}$ (or
 27 $\mu_{\ln(-\beta_k)}$ and $\sigma_{\ln(-\beta_k)}$ if we work with a negative exponential) denote the mean and
 28 standard deviation of the underlying Normal distribution, where ξ_{nk} follows a standard
 29 Normal distribution across respondents for attribute k . It can be observed that as

1 $e^{\tau_k \alpha_{nk}}$ itself follows a Lognormal distribution, β_{nk} does too as it is formed by a product
 2 of Lognormals.

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

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

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

6 **2.4. Measurement model components**

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

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

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

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

$$W_{jnm|c} = -\lambda_{j|c} B_{jnm|c}, \quad (7)$$

24 thus accounting for scale difference between the “best” and the “worst” stage and al-
 25 lowing this difference to be attribute-specific, while still assuming that the driving fac-
 26 tors of making an attribute (level) important/attractive or unimportant/unattractive
 27 are the same across the two stages. Hence this specification is different from the original

⁵The 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.

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

4 Due to the experimental design, the choice set varies over respondents and tasks,
 5 and this thus affects what is possible for a respondent to select as the combination
 6 of best and worst attributes or attribute levels in a given scenario. We use $\mathbb{D}_{nm|c}$ to
 7 define the set containing all the available items presented to respondent n in BWS
 8 task m and type of BWS data c . The items in $\mathbb{D}_{nm|c}$ allow the formation of the set
 9 $\mathbb{S}_{nm|c}$ containing all the possible best-worst pairs of the available attributes or attribute
 10 levels, respectively. Similar to other MNL models with a RUM assumption, the best-
 11 worst choice probabilities of respondent n selecting h as the best and r as the worst
 12 ($h, r \in \mathbb{D}_{nm|c}, r \neq h, (h, r) \in \mathbb{S}_{nm|c}$) in BWS task m can then be written as:

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

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

14 2.4.1. BWS1 data

15 In the BWS1 setting, we work with attributes rather than attribute levels. The “util-
 16 ity” function is specified to represent the weight placed on an attribute k by respondent
 17 n in task m in decision-making. Thus we have a single “utility” for a given attribute
 18 k to be “best” attribute, which is given by:⁶

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

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

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

⁶In 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.

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

3 2.4.2. BWS2 data

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

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

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

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

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

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

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

1 wider gap in terms of attractiveness between a higher level and the lowest (base) level
 2 for that attribute than others do.

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

6 **2.5. Log-likelihood**

7 The unconditional probability of observing the sequence of stated choices y_n and best-
 8 worst responses $(b, w)_n$ can be expressed as the integral of the multiplication of the
 9 conditional stated choice probabilities and the conditional best-worst choice probabil-
 10 ities over the distribution of η_n , the random component of the latent variables α_n ,
 11 and over the distribution of ξ_n , the random component of the unobserved preference
 12 heterogeneity irrelevant from α_n , such that the log-likelihood is given by:

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

13 where T_n , $M_{n|1}$ and $M_{n|2}$ give the total numbers of the SC tasks, the BWS1 tasks, and
 14 the BWS2 tasks shown to respondent n . Meanwhile, choice observations y_{nt} , $(b, w)_{nm|1}$,
 15 $(b, w)_{nm|2}$ refer to the chosen alternative in a SC task, the chosen best-worst pair of
 16 attributes in a BWS1 task, and the best-worst pair of attribute levels selected in a
 17 BWS2 task, respectively. Since the resulting LL does not have closed-form expression,
 18 the value of the log-likelihood needs to be approximated through simulation (Train
 19 2009).

20 **2.6. Hypothesis**

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

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

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

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

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

23 **3. Case study: Survey and data**

24 **3.1. Survey background**

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

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

⁷A 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.

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

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

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

15 3.2. SC tasks

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

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

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

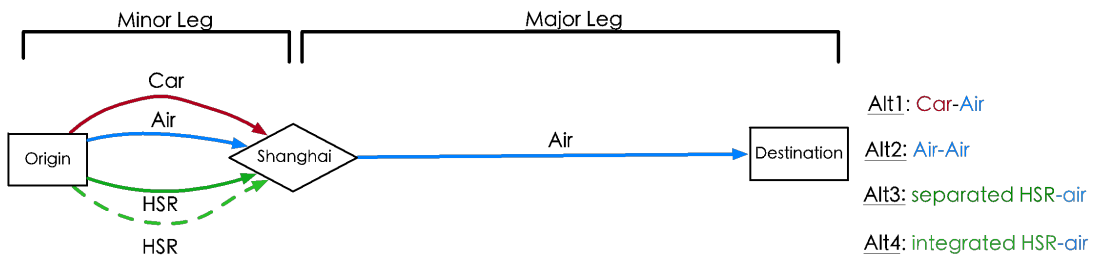


Figure 2.: Illustration of choice scenarios in the SC survey.

29 The SC survey was generated through a *D*-efficient design (Rose and Bliemer 2007)

1 in Ngene (Metrics 2012). Each respondent was presented with 8 SC tasks in a ran-
2 domised order, giving a total of 984 stated choice observations. Fig. 3 shows an example
3 of the SC tasks. A total of 7 attributes were incorporated, including minor time, con-
4 nection time, transfer time, delay protection, ticket integration, luggage integration
5 and travel cost. Minor time gives the time spent on the minor leg; transfer time de-
6 notes the time spent on transferring between the minor leg and the major leg;⁸ and
7 connection time means the time spent on waiting and going through various proce-
8 dures (e.g. security check-in, luggage check-in) at the departure airport of the major
9 leg. Travel cost gives the total expenditure for the journey, and delay protection in-
10 dicates to what extent a respondent would be compensated in case of delay on the
11 minor leg. Ticket integration and luggage integration are two attributes describing the
12 extent of integration of the ticketing systems and luggage-handling systems between
13 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	-	<ul style="list-style-type: none"> • Book together • Fixed-time flight on minor leg • Easy collection 	<ul style="list-style-type: none"> • Book separately • Fixed-time train on minor leg • No easy collection 	<ul style="list-style-type: none"> • Book together • Fixed-time train on minor leg • Easy collection
Security check and luggage integration	-	<ul style="list-style-type: none"> • Two security checks • No integrated luggage handling system 	<ul style="list-style-type: none"> • Two security checks • No integrated luggage handling system 	<ul style="list-style-type: none"> • One security check • Integrated luggage handling system available

Figure 3.: Example of SC tasks.

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

⁸Transfer 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.

1 **3.3. BWS Case 1 tasks**

2 The BWS1 section required respondents to choose the attributes that they weighted
 3 the most and the least in each task. A balanced incomplete block design (BIBD)
 4 was adopted to generate the BWS1 experiment which could ensure each attribute
 5 occurred the same number of times and co-occurred with any other attribute the same
 6 number of times across all the choice tasks (Louviere, Flynn, and Marley 2015). In our
 7 survey, 7 attributes were assigned into 7 randomly-displayed BWS1 tasks, each with
 8 4 attributes. Consequently, each attribute was shown to each respondent 4 times and
 9 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
<input type="checkbox"/>	Minor time	<input type="checkbox"/>
<input type="checkbox"/>	Delay protection	<input type="checkbox"/>
<input type="checkbox"/>	Connection time	<input type="checkbox"/>
<input type="checkbox"/>	Travel cost	<input type="checkbox"/>

Figure 4.: Example of BWS1 tasks.

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

⁹Simple 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

1 3.4. BWS Case 2 tasks

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

14 Fig. 5 gives an example of the BWS2 tasks, where different levels across different
15 attributes were evaluated on a common scale rather than being compared within an
16 attribute, such that a respondent might prefer “having 50% off on a flight change”
17 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
<input type="checkbox"/>	Connection time: 2.5h	<input type="checkbox"/>
<input type="checkbox"/>	50% off on changing flight	<input type="checkbox"/>
<input type="checkbox"/>	Book together, fixed-time train on the minor leg and easy collection	<input type="checkbox"/>
<input type="checkbox"/>	Integrated luggage-handling and one security check	<input type="checkbox"/>

Figure 5.: Example of BWS2 tasks.

¹⁰The 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

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

4 It should be noted that each item was not necessarily presented to all of the 123
 5 respondents and did not occur with a same frequency. Thus, we calculate analytical B-
 6 W scores¹¹ to show relative attractiveness of the attribute levels among the sample. As
 7 shown in Table 3, we can see an increase in the analytical B-W scores as the level goes
 8 up for delay protection and luggage integration. However, for ticket integration, the
 9 scores are generally low and close to each other, indicating that the three levels of ticket
 10 integration were almost equally attractive to the respondents. One interesting thing is
 11 that connection time appears to be generally considered less attractive, regardless of
 12 which actual value it takes. This is understandable as connection time was considered
 13 as the most important factor in the BWS1 tasks, so that the respondents felt all the
 14 values of connection time presented in the BWS2 tasks to be unattractive.

15 The scores are used for descriptive analysis for better understanding the BWS1 and
 16 BWS2 data. All in all, we wish to study the correlation across the different datasets.
 17 The B-W scores themselves do not allow us to do so because we can only calculate the
 18 scores for BWS1 and BWS2 data independently, regardless of the calculation method
 19 we adopt. We need the joint model to simultaneously estimate on SC, BWS1 and
 20 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

¹¹Analytical 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).

1 4. Case study: Model estimation

2 4.1. Model specification

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

14 4.1.1. Structural equations

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

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

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

¹²For 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.

1 *4.1.2. Choice model for SC data*

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

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

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

11 such that $\beta_{n,MT}$, $\beta_{n,CT}$ and $\beta_{n,TC}$ measure the marginal utilities, while $\beta_{n,tran45\&90min}$,
12 $\beta_{n,delay1\&2}$, and $\beta_{n,lugg1\&2}$ give the relative utility against the corresponding base lev-
13 els, which are $tran0min$, $delay0$, and $lugg0$ in respective. The higher two levels for each
14 are merged for estimation in our final specification as they are found not significantly
15 different from each other. The final specification excludes the attribute of ticket inte-
16 gration from the utility function for the SC data, as it is found to contribute little to
17 the utility functions. However, ticket integration is still used in the measurement mod-
18 els. Finally, parameters of $\kappa_{DP,male}$, $\kappa_{TC,reimbursed}$ and τ_{DP} are set to zero in the final
19 specification as they were insignificant. Besides, although we have found suitable socio
20 to explain transfer time (i.e. $Z_{age>45}$), the model with the indirect impact of $Z_{age>45}$
21 becomes insignificant once the direct impact is added. Hence, in the final specification,
22 we drop the indirect impact by fixing $\omega_{TT,age>45} = 0$ and keep the direct impact of
23 age on transfer time by estimating $\kappa_{TT,age>45}$.

24 *4.1.3. Measurement models for BWS1 data and BWS2 data*

25 For the BWS1 data, all the 7 attributes shown in the SC survey are examined, i.e. mi-
26 nor time, connection time, transfer time, delay protection, ticket integration, luggage
27 integration and travel cost. Minor time acts as the base, with relevant parameters
28 $\delta_{MT|1}$ and $\zeta_{MT|1}$ normalised to 0. For the BWS2 data, connection time, delay pro-

1 tection, ticket integration and luggage integration are the four attributes of interest.
 2 Connection time is treated as a continuous variable and $x_{CT, nm|2}$ can take the value
 3 of 150min, 180min, 210min, 270min or 330min. The remaining three attributes are
 4 regarded as categorical variables, with level $delay0$, $tick1$ and $lugg0$ being the lowest
 5 (base) levels for delay protection, ticket integration and luggage integration in respec-
 6 tive. The attribute level $delay0$ is selected as the base in the measurement model for
 7 BWS2 data, with the baseline attractiveness $\phi_{delay0|2}$ fixed to 0 for normalisation.

8 4.2. Estimation results

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

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

26 The MMNL model and the ICLV model show similar preference patterns towards
 27 attributes. As shown in the upper part of Table 4, the most negative δ_{ca} implies that
 28 the car-air alternative is the least preferred option, all else being equal, whereas the
 29 air-air alternative (δ_{aa}) and the separated HSR-air alternative (δ_{sha}) are both slightly
 30 less preferred compared to the base alternative, i.e. the integrated HSR-air mode. Since
 31 Lognormal distributions are used, the more negative the underlying mean parameter
 32 $\mu_{\ln|\beta_k|}$ is, the smaller in magnitude the median of marginal utility is, which translates
 33 into a lower sensitivity to that attribute in the SC tasks. As to the standard deviations
 34 $\sigma_{\ln|\beta_k|}$, both models detect statistically significant random heterogeneity in sensitivities
 35 to all of the attributes. Regarding the direct impacts of socio-demographics in the
 36 utility functions, we can see from both models that $\kappa_{TT, age>45}$ is significant at the 95%
 37 confidence interval, suggesting that older respondents are more sensitive to transfer

¹³USD/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

Log likelihood	MMNL		ICLV	
	LL: -1057.396		LL (total): -4445.399 LL (SC): -1060.453	
	est	t-rat(0)	est	t-rat(0)
δ_{ca}	-3.210	-7.49	-3.081	-6.91
δ_{aa}	-0.411	-1.73	-0.439	-2.04
δ_{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	-
τ_{CT}			0.233	2.37
τ_{TT}			0.335	2.59
τ_{DP}			0.000	-
τ_{LI}			0.701	4.49
τ_{TC}			0.334	1.21

1 time and dislike long transfer time more than young people do. Meanwhile, although
2 $\kappa_{LI,age>45}$ in the MMNL model is only significant at the 80% confidence interval, we
3 can still infer from $\kappa_{LI,age>45}$ in the ICLV model, which is significant at the 95%
4 confidence interval, that older passengers can derive higher utility from better luggage
5 integration than young people do.

6 In the left part of Table 5, the constant $\delta_{|1}$ represents the mean of the weight to the
7 associated attribute among the sample in the BWS1 data. It could be noticed that,
8 with minor time normalised to 0, connection time, delay protection and transfer time
9 are positioned at the higher end of the underlying weighting scale, followed by travel
10 cost and luggage integration. Regarding the scalars in the worst choice stage shown in
11 the down left of Table 5, $\lambda_{CT|1}$ (t-rat(1)=-4.27) is the only one which is significantly
12 different from 1, suggesting that scaling difference between the worst choice stage and
13 the best choice stage only exists for the attribute of connection time. Since $\lambda_{CT|1}$ is
14 much lower than 1, it suggests that the model has less noise in explaining the choices
15 in the best choice stage than in the worst choice stage for the attribute of connection
16 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	$\lambda_{TC 2}$	-	-	-

17 The right part of Table 5 shows estimates for the baseline attractiveness of each
18 attribute level in the BWS2 data. Focusing on $\phi_{|2}$, it can be inferred that compared to
19 ticket integration, delay protection and luggage integration are associated with overall
20 larger steps in attractiveness when moving from a poorer level to a better level, which
21 implies that respondents might be indifferent to variations in ticket integration. This is

1 in line with the discoveries in the SC data and the BWS1 data as well as the preliminary
2 findings in the normalised B-W scores in the BWS2 data. As to the attribute-specific
3 scalars shown in the down right of Table 5, only ticket integration $\lambda_{TI|2}$ (t-rat(1)=-
4 2.42) and luggage integration $\lambda_{LI|2}$ (t-rat(1)=-2.13) are significantly different from 1.
5 Being smaller than 1, $\lambda_{TI|2}$ and $\lambda_{LI|2}$ suggest stronger random error in the worst choice
6 stage for these two attributes than in the best choice stage.

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

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

40 Combining the estimates ω 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)		
ω_{MT}	-	-	τ_{MT}	-	-	$\zeta_{MT 1}$	0 (base)	-	$\zeta_{MT 2}$	-	-
ω_{CT}	-	-	τ_{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	-	τ_{TT}	0.335	2.59	$\zeta_{TT 1}$	1.211	4.50	$\zeta_{TT 2}$	-	-
$\omega_{DP,male}$	-0.863	-2.71	τ_{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	τ_{TI}	-	-	$\zeta_{TI 1}$	1.683	4.34	$\zeta_{TI 2}$	0.371	3.94
$\omega_{LI,age>45}$	1.191	2.66	τ_{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	τ_{TC}	0.334	1.21	$\zeta_{TC 1}$	2.210	5.66	$\zeta_{TC 2}$	-	-

1 latent attribute importance, the positive $\omega_{TI,age>35}$ and $\omega_{LI,age>45}$ and the negative
2 $\omega_{TC,reimbursed}$ show that older people think ticket integration and luggage integration
3 to be of greater importance than young people do, while passengers who get reimbursed
4 perceive lower importance for travel cost than those who need to pay for the travel
5 on their own. The negative and significant $\omega_{DP,male}$ suggests that male passengers
6 find delay protection less important than female passengers do. Parameter $\omega_{TT,age>45}$
7 are fixed to 0 and not estimated in the final specification because of its very low
8 significance. We can further look back into Table 4, where $\kappa_{TT,age>45}$ and $\kappa_{LI,age>45}$
9 are the only two statistically significant κ parameters. We can therefore deduce that
10 respondents' age mainly plays an independently direct role in scaling the marginal
11 utility of transfer time, whereas age affects the marginal utility of luggage integration
12 both directly and indirectly via the latent variable. The remaining socio-demographic
13 characteristics involved in ω influence stated choice behaviour mainly through the
14 latent variables of attribute importance.

15 Finally, we shed some light on willingness-to-pay (WTP) in the SC data with and
16 without the additional information gained from the BWS1 and BWS2 data in Table
17 7. We first calculated the distributions of marginal utilities for all the attributes,
18 taking into account of the roles of latent attribute importance and socio-demographic
19 characteristics in the ICLV model and the role of socio-demographic characteristics in
20 the reduced form MMNL model, i.e. marginal utilities β_{nk} are given by $e^{\tau_k \alpha_{nk}} e^{\kappa_k Z_n} \beta_{nk}^*$
21 in the ICLV model and by $e^{\kappa_k Z_n} \beta_{nk}^*$ in the MMNL model, where $\beta_{nk}^* = e^{\mu_{ln\beta_k} + \sigma_{ln\beta_k} \cdot \xi_{nk}}$.
22 We then calculated the ratio against the marginal utility of travel cost for each of the
23 remaining attributes for each draw, which is taken from the distributions of marginal
24 utilities used in the estimation procedure, enabling us to obtain the WTP distributions
25 for all the attributes except for travel cost through simulation (Hensher and Greene
26 2003; Sillano and de Dios Ortúzar 2005; Daly, Hess, and Train 2012).

27 We see some differences between the two models here, where we would argue that
28 the ICLV findings are more realistic especially for transfer time. Indeed, in the ICLV
29 model, going from a transfer time of 45 or 90 minutes to a seamless transfer has the
30 same benefit as a reduction in connection time by 81.6 minutes at the mean. In the
31 MMNL model, this would be 122.58 minutes, which seems unrealistic if we assume

1 that transfer time should at best be as important as connection time. In addition,
2 the standard deviations of the three categorical attributes, i.e. transfer time, delay
3 protection, and luggage integration are relatively large in both models. This can be
4 mainly attributed to the long tails of the Lognormal distributed WTP distributions
5 as the marginal utilities for all the attributes follow Lognormal distributions. Hence,
6 apart from regular statistics of mean and standard deviation, we also show the median
7 and interquartile range of each WTP distribution. We can see an overall reduction in
8 the median values, and a decrease in the interquartile range for all the attributes except
9 for minor time when we move from the MMNL model to the ICLV model. This means
10 that the spread of the distribution is smaller and the values are more squeezed to
11 the median for the ICLV model.

1

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

models	attributes	sensitivities β		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	-	-	-	-	-	-	-	-

2

1 5. Conclusions

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

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

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

33 Overall, our joint model shows that attribute importance can link the SC, BWS1 and
34 BWS2 data, indicating the benefit of improving behavioural explanation by combining
35 the BWS data with SC data. We found a high level of consistency with respect to the
36 impact of the underlying perceived *attribute importance* on decision-making in different
37 tasks is significantly demonstrated. The estimation results imply that an increase in
38 attribute importance results in a stronger sensitivity to that attribute in the SC tasks,
39 more overall weight to that attribute in the BWS1 tasks, and also wider attractiveness

1 gaps between levels for that attribute in the BWS2 tasks. This is particularly true for
2 non-cost attributes, including connection time, transfer time and luggage integration
3 in our case. We have not found similar consistency for cost-relevant attributes, i.e.
4 delay protection and travel cost, as the corresponding latent variables only impose
5 significant impacts in the BWS1/2 data but not in the SC data. That is, we have not
6 discovered a one-to-one relationship between different survey methods. As such, there
7 remain some differences in how attribute importance is evaluated between SC, BWS1
8 and BWS2 data. We therefore think treating different survey methods as equivalent
9 and interchangeable - for example using BWS1 method to determine which attributes
10 to include in SC survey - can be risky.

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

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

1 research contexts.

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

12 **Acknowledgements**

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

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

α	Matrix, giving the latent <i>attribute importance</i> of each attribute perceived by each respondent.
α_n	Vector, giving the latent <i>attribute importance</i> of each attribute perceived by respondent n .
α_{nk}	Scalar, giving the latent <i>attribute importance</i> of attribute k perceived by respondent n .
β	Matrix, describing the marginal utility of each attribute for each respondent.
β_n	Vector, describing the marginal utility of each attribute for respondent n .
β_k	Vector, describing the marginal utility of attribute k for each respondent.
β_{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.
2 $B_{qnm c}$	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) .
δ_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).
η_{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).

κ	Matrix, describing the impact of each socioeconomic variable on each attribute's marginal utility.
κ_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 β_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 .
ω	Matrix, describing the impact of each socio-demographic variable on each attribute's corresponding latent <i>attribute importance</i> .
ω_k	Vector, measuring the impact of each socio-demographic variable on the latent <i>attribute importance</i> 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 variable).
1	
$\sigma_{\ln\beta_k}$	Scalar, capturing the standard deviation of the underlying Normal distribution for β_k .
τ	Vector, describing the impact of each latent <i>attribute importance</i> on the corresponding attribute's marginal utility in the SC component.
τ_k	Scalar, describing how the marginal utility of attribute k is affected by the corresponding <i>attribute importance</i> 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 .
ε_{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 .
ξ_{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 <i>attribute importance</i> 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 <i>attribute importance</i> in the BWS1 component.
¹ $\zeta_{ 2}$	Vector, describing the impact of each latent <i>attribute importance</i> 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 <i>attribute importance</i> 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 .

1 References

- 2 Auger, Pat, Timothy M Devinney, and Jordan J Louviere. 2007. "Using best-worst scaling
3 methodology to investigate consumer ethical beliefs across countries." *Journal of Business*
4 *Ethics* 70 (3): 299–326.
- 5 Balbontin, Camila, J de D Ortúzar, and JD Swait. 2015. "A joint best-worst scaling and
6 stated choice model considering observed and unobserved heterogeneity: An application to
7 residential location choice." *Journal of choice modelling* 16: 1–14.
- 8 Beck, Matthew J, and John M Rose. 2016. "The best of times and the worst of times: A
9 new best-worst measure of attitudes toward public transport experiences." *Transportation*
10 *Research Part A: Policy and Practice* 86: 108–123.
- 11 Beck, Matthew J, John M Rose, and Stephen P Greaves. 2017. "I cant believe your attitude: a
12 joint estimation of best worst attitudes and electric vehicle choice." *Transportation* 44 (4):
13 753–772.
- 14 Ben-Akiva, Moshe, Joan Walker, Adriana T Bernardino, Dinesh A Gopinath, Taka Morikawa,
15 and Amalia Polydoropoulou. 2002. "Integration of choice and latent variable models." *Per-*
16 *petual motion: Travel behaviour research opportunities and application challenges* 431–470.
- 17 Campbell, Danny, David A Hensher, and Riccardo Scarpa. 2011. "Non-attendance to attributes
18 in environmental choice analysis: a latent class specification." *Journal of environmental*
19 *planning and management* 54 (8): 1061–1076.
- 20 Carlsson, Fredrik. 2003. "The demand for intercity public transport: the case of business
21 passengers." *Applied Economics* 35 (1): 41–50.
- 22 Coast, Joanna, Terry N Flynn, Chris Salisbury, Jordan Louviere, and Tim J Peters. 2006.
23 "Maximising responses to discrete choice experiments." *Applied health economics and health*
24 *policy* 5 (4): 249–260.
- 25 Collins, Andrew T, Michiel CJ Bliemer, and John M Rose. 2014. "Constrained stated choice
26 experimental designs." .
- 27 Daly, Andrew, Stephane Hess, and Kenneth Train. 2012. "Assuring finite moments for will-
28 ingness to pay in random coefficient models." *Transportation* 39 (1): 19–31.
- 29 Dumont, Jeffrey, Marek Giergiczny, and Stephane Hess. 2015. "Individual level models vs.
30 sample level models: contrasts and mutual benefits." *Transportmetrica A: Transport Science*
31 11 (6): 465–483.
- 32 Dyachenko, Tatiana, Rebecca Walker Reczek, and Greg M Allenby. 2014. "Models of sequential
33 evaluation in best-worst choice tasks." *Marketing Science* 33 (6): 828–848.
- 34 Finn, Adam, and Jordan J Louviere. 1992. "Determining the appropriate response to evidence
35 of public concern: the case of food safety." *Journal of Public Policy & Marketing* 12–25.
- 36 Giergiczny, Marek, Thijs Dekker, Stephane Hess, and P Chintakayala. 2017. "Testing the
37 stability of utility parameters in repeated best, repeated best-worst and one-off best-worst
38 studies." *European Journal of Transport and Infrastructure Research* 17 (4): 457–476.
- 39 Hawkins, Guy E, Towhidul Islam, and AAJ Marley. 2018. "Like it or not, you are using one
40 value representation." *Decision* .
- 41 Hensher, David A. 2006. "How do respondents process stated choice experiments? Attribute
42 consideration under varying information load." *Journal of applied econometrics* 21 (6): 861–

- 1 878.
- 2 Hensher, David A, and William H Greene. 2003. "The mixed logit model: the state of practice."
3 *Transportation* 30 (2): 133–176.
- 4 Hensher, David A, Corinne Mulley, and John M Rose. 2015. "Understanding the relationship
5 between voting preferences for public transport and perceptions and preferences for bus
6 rapid transit versus light rail." *Journal of Transport Economics and Policy (JTEP)* 49 (2):
7 236–260.
- 8 Hensher, David A, John Rose, and William H Greene. 2005. "The implications on willingness
9 to pay of respondents ignoring specific attributes." *Transportation* 32 (3): 203–222.
- 10 Hensher, David A, and John M Rose. 2009. "Simplifying choice through attribute preservation
11 or non-attendance: implications for willingness to pay." *Transportation Research Part E:
12 Logistics and Transportation Review* 45 (4): 583–590.
- 13 Hess, Stephane, and David A Hensher. 2010. "Using conditioning on observed choices to re-
14 trieve individual-specific attribute processing strategies." *Transportation Research Part B:
15 Methodological* 44 (6): 781–790.
- 16 Hess, Stephane, and David A Hensher. 2013. "Making use of respondent reported process-
17 ing information to understand attribute importance: a latent variable scaling approach."
18 *Transportation* 40 (2): 397–412.
- 19 Hess, Stephane, and David Palma. 2019. "Apollo: A flexible, powerful and customisable free-
20 ware package for choice model estimation and application." *Journal of choice modelling* 32:
21 100170.
- 22 Hess, Stephane, Amanda Stathopoulos, Danny Campbell, Vikki O'Neill, and Sebastian Caus-
23 sade. 2013. "Its not that I dont care, I just dont care very much: confounding between
24 attribute non-attendance and taste heterogeneity." *Transportation* 40 (3): 583–607.
- 25 Hess, Stephane, Kenneth E Train, and John W Polak. 2006. "On the use of a modified latin
26 hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle
27 choice." *Transportation Research Part B: Methodological* 40 (2): 147–163.
- 28 Lipovetsky, Stan, and Michael Conklin. 2014. "Best-worst scaling in analytical closed-form
29 solution." *Journal of choice modelling* 10: 60–68.
- 30 Louviere, Jordan J, Terry N Flynn, and Anthony AJ Marley. 2015. *Best-worst scaling: Theory,
31 methods and applications*. Cambridge University Press.
- 32 Louviere, Jordan J, and Towhidul Islam. 2008. "A comparison of importance weights and
33 willingness-to-pay measures derived from choice-based conjoint, constant sum scales and
34 best–worst scaling." *Journal of Business Research* 61 (9): 903–911.
- 35 Marley, AAJ, T Islam, and GE Hawkins. 2016. "A formal and empirical comparison of two
36 score measures for best–worst scaling." *Journal of choice modelling* 21: 15–24.
- 37 Marley, Anthony AJ, and Jordan J Louviere. 2005. "Some probabilistic models of best, worst,
38 and best–worst choices." *Journal of Mathematical Psychology* 49 (6): 464–480.
- 39 Marley, Anthony Alfred John, Terry N Flynn, and Jordan Joseph Louviere. 2008. "Probabilistic
40 models of set-dependent and attribute-level best–worst choice." *Journal of Mathematical
41 Psychology* 52 (5): 281–296.
- 42 Marti, Joachim. 2012. "A best–worst scaling survey of adolescents' level of concern for health
43 and non-health consequences of smoking." *Social Science & Medicine* 75 (1): 87–97.

- 1 McFadden, Daniel. 2001. "Economic choices." *American economic review* 91 (3): 351–378.
- 2 McFadden, Daniel, et al. 1973. "Conditional logit analysis of qualitative choice behavior." .
- 3 Metrics, Choice. 2012. "Ngene 1.1. 1 User Manual & Reference Guide." *Sydney, Australia:*
4 *ChoiceMetrics* .
- 5 Mueller, Simone, Larry Lockshin, and Jordan J Louviere. 2010. "What you see may not be
6 what you get: Asking consumers what matters may not reflect what they choose." *Marketing*
7 *Letters* 21 (4): 335–350.
- 8 Petrolia, Daniel R, Matthew G Interis, and Joonghyun Hwang. 2018. "Single-Choice, Repeated-
9 Choice, and Best-Worst Scaling Elicitation Formats: Do Results Differ and by How Much?"
10 *Environmental and Resource Economics* 69 (2): 365–393.
- 11 Potoglou, Dimitris, Peter Burge, Terry Flynn, Ann Netten, Juliette Malley, Julien Forder,
12 and John E Brazier. 2011. "Best–worst scaling vs. discrete choice experiments: an empirical
13 comparison using social care data." *Social science & medicine* 72 (10): 1717–1727.
- 14 Pullman, Madeleine E, Kimberly J Dodson, and William L Moore. 1999. "A comparison of
15 conjoint methods when there are many attributes." *Marketing Letters* 10 (2): 125–138.
- 16 Rose, John M. 2014. *Interpreting discrete choice models based on best-worst data: a matter of*
17 *framing*. Technical Report.
- 18 Rose, John M, and Michiel CJ Bliemer. 2007. "Stated preference experimental design strate-
19 gies." In *Handbook of Transport Modelling: 2nd Edition*, 151–180. Emerald Group Publishing
20 Limited.
- 21 Sillano, Mauricio, and Juan de Dios Ortúzar. 2005. "Willingness-to-pay estimation with mixed
22 logit models: some new evidence." *Environment and Planning A* 37 (3): 525–550.
- 23 Song, Fangqing, Stephane Hess, and Thijs Dekker. 2018. "Accounting for the impact of variety-
24 seeking: Theory and application to HSR-air intermodality in China." *Journal of Air Trans-*
25 *port Management* 69: 99–111.
- 26 Train, Kenneth E. 2009. *Discrete choice methods with simulation*. Cambridge university press.
- 27 Vij, Akshay, and Joan L Walker. 2016. "How, when and why integrated choice and latent
28 variable models are latently useful." *Transportation Research Part B: Methodological* 90:
29 192–217.