COMPARING AND COMBINING BEST-WORST SCALING AND STATED CHOICE DATA TO UNDERSTAND ATTRIBUTE IMPORTANCE IN MODE CHOICE BEHAVIOUR

Fangqing Song, Corresponding Author
Institute for Transport Studies, University of Leeds
Choice Modelling Centre, University of Leeds
LS2 9JT, Leeds, UK
Tel: (+44) 07752389924; Email: tsfs@leeds.ac.uk

Stephane Hess
Institute for Transport Studies, University of Leeds
Choice Modelling Centre, University of Leeds
LS2 9JT, Leeds, UK
Tel: (+44) 01133436611; Email: S.Hess@its.leeds.ac.uk

Thijs Dekker
Institute for Transport Studies, University of Leeds
Choice Modelling Centre, University of Leeds
LS2 9JT, Leeds, UK
Tel: (+44) 01133435334; Email: T.Dekker@leeds.ac.uk

ABSTRACT

A large share of travel behaviour research is concerned with accommodating heterogeneity in preferences across individual travellers. Much of this work is conducted using random coefficients models such as Mixed Logit, estimated on choice data either from revealed preference or stated choice survey. In contrast, other areas of research have increasingly made use of best-worst scaling (BWS) exercises in which respondents assign importance to different attributes outside a multi-alternative context. The present paper contrasts these two approaches in the context of a survey of HSR (high speed rail)-air intermodality in China. Using different approaches, including descriptive analysis, Bayesian posteriors and hybrid choice models, we find a certain level of correspondence between the behaviour in the stated choice scenarios and the responses from the BWS exercises. This is especially strong for qualitative attributes but also travel cost.

Keywords: Attribute importance, Best-worst scaling, Hierarchical Bayesian estimation, Hybrid choice model

INTRODUCTION

Accommodating heterogeneity in preferences across individual decision makers has become one of the most active areas of research in choice modelling and travel behaviour analysis in particular. Much of the work makes use of Mixed Logit models and seeks to acknowledge the fact that some individuals have higher sensitivities for a specific subset of attributes, while others individuals care more about a different subset. Other work has looked at the notion that some individuals, especially in stated choice (SC) surveys, may completely ignore certain attributes, a notion however challenged by other work showing that these individuals may simply care less about these attributes.

While the majority of the work in travel behaviour research relies on either revealed preference or stated choice data, research especially outside of transport has increasingly made use of best-worst scaling (BWS, or maxdiff) approaches to determine which attributes matter more for a given person, and which attributes matter less. A key question then arises whether the way in which respondents rank the attributes in importance in a BWS exercise is consistent with how those same attributes influence the choices in a multi-alternative setting.

Our work makes use of a survey in which all respondents not only completed a SC survey but also provided answers to best-worst (BW) scoring tasks that allow us to elicit a full ordering of importance for the different attributes. We contrast the results from the two survey components in different ways, including descriptively looking at the relationship between the BW scores and the choice strategies employed in the data, contrasting the findings from Bayesian estimation of individual-level coefficient distributions from the SC survey with the BW scores, and finally estimating a hybrid choice models making use of the BW scores as indicators of attribute importance, following the approach in (*1*) which relied on stated attribute non-attendance. In contrast with that work, a richer pattern of data is available here, given that for each attribute we have score, rather than just a 0-1 response.

What is novel in our current research is combining the traditionally used SC tasks with the increasingly popularised BWS approach in detecting the existence of heterogeneity across respondents. To the author's knowledge, though some studies have already compared the estimates from traditional SC surveys with those gained from BW tasks (*2, 3, 4, 5*), and some other studies have modelled BW data through discrete choice modelling techniques, no attempts have been made to combine SC data and BW data within a single model framework. We take advantages of linking these two types of data in improving our understanding of choice behaviour, including attribute importance. In this sense, this research aims at filling this gap while providing more empirical evidence in investigating attribute importance.

The remainder of this paper is organised as follows. We first present the data before three separate sections which look at the different analyses carried out on the data. We then present some conclusions.

DATA

Our work makes use of data from a survey on HSR-air intermodality conducted in Shanghai, China. The survey is framed around a situation where: 1) a passenger is travelling from a domestic origin O to an overseas destination D; 2) direct flights from O to D are unavailable; 3) a passenger from O to D needs to travel via Shanghai; and 4) a passenger can only travel by air between Shanghai and D. Four alternatives were presented to respondents, namely car-air, air-air, separated HSR-air and integrated HSR-air. We denote the first leg between O and Shanghai as the "minor leg" on which various modes are available, and the second leg between Shanghai and D as the "major leg" where air is the only option. Car-air means using car on the minor leg and using flight on the major leg; air-air means connecting flights; separated HSR-air refers to the traditional travel which need purchasing air and HSR tickets separately; integrated HSR-air refers to the new HSR-air intermodal service. The data is collected at Pudong International airport in January 2017 through face-to-face interview with online questionnaires which include a stated choice (SC) component, and a best-worst (BW) component and some other tasks.

In total, 123 respondents completed the whole questionnaire. In the SC component, each respondent was presented with 8 SC tasks in a randomised order, each with 4 alternatives mentioned above, giving a total of 984 choice observations. The SC survey was generated through D-efficient design (*6*) in Ngene (*7*). A total of 7 attributes were used, including travel time on the minor leg, connection time, transfer time, protection in case of delay on minor leg, integration of ticketing system, security check and luggage integration[1], and total travel cost. The transfer time means the moving time between the two legs which in particular takes a value of 0 for a seamless transfer at an intermodal hub. Connection time refers to the time spent on waiting and going through procedures. The sum of transfer time and connection time gives the time intervals between the departure time of the major leg and the arrival time of the minor leg. Transfer time is fixed to zero for car, while it can also take a value of zero for any of the other alternatives. For car, connection time is fixed to the minimum pre-departure arrival time of 90 minutes. Delay protection, presented in 3 different levels gives information on how the respondent would be compensated in case of delay on the minor leg resulting in missing the flight on the major leg. Ticket integration describes the integration level of air and HSR ticketing systems, with 4 different levels available. Security check and luggage integration refers to how many security checks and luggage check-in are required throughout the travel, with 3 different levels. These two final attributes (ticket integration and security check and luggage integration) do not apply for the car-air alternative and are kept at the lowest level for the segregated HSR-air alternative. Figure 1 shows an example of SC task.

BWS section came after SC component, and required respondents to choose the best one and the worst one from a choice set, where "best" and "worst" can be replaced with other proper words representing the two extremes of the "continuum" according to research background. BWS approach includes three different types (*8*). Case 1, also called object case, compares between different attributes themselves without considering their levels; Case 2, also known as profile case, compares between different attribute levels within a profile that discribe an alternative; Case 3 also named as multi-profile case, compares between different profiles which could be equivalent to discrete choice experiment. We adopt BW case 1 to measure the seven attributes which describe the integrated HSR-air alternative in SC tasks, given its advantages over the traditionally used ranking or rating methods. It is easier for respondents to select the two extreme options in a relatively smaller choice set in a BW task especially when respondents need to order many attributes, which would otherwise be difficult in ranking tasks (*9*). Besides, BWS can avoid the risk of lacking discrimination of the data which might arise in rating tasks as respondents do not need to make serious trade-off among different items. In addition, though BW tasks may be more tedious than the other two for respondents, the collected data can provide much more readily understandable and managerially meaningful results to analysts (*10*).

[1] In the remainder of the paper, this attribute is simplified as "luggage integration" for reference

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

FIGURE 1 Example of SC task.[a]
[a] CNY/ USD was around 0.145 during the survey period

A balanced incomplete block design (BIBD) was adopted to generate BW tasks as it is the most prevalent design method that can ensure each item occurring the same often and co-occurring with any other item the same often across all the choice tasks which are of the same size (*8*). In our survey, 7 BW tasks measuring on 7 attributes were presented to respondents in a randomise order, each with 4 attributes among which respondents were required to select the most important and least important attributes they considered during the SC component. Consequently, each attribute was shown to respondents 4 times and each pair of attributes also occurred 2 times. In this way, our respondents did not confront with some attributes occurring more times or came across different size of choice tasks in the survey, which may lead to biasedness otherwise.

ANALYSIS I: DESCRIPTIVE COMPARISON BETWEEN BW SCORES AND CHOICE BEHAVIOUR

Method

A simple way to evaluate the data from BW surveys is to calculate the aggregated best-minus-worst (B-W) scores by subtracting how many times an item is chosen as the worst from how many times it is chosen as the best across all tasks and across respondents (*8*). B-W scores can measure a continuum of interest, and in our case, higher B-W score means greater importance or preference. The B-W score of each attribute across all BW tasks and across all respondents indicates how important the attribute is or how the attribute is preferred in the sample. Although no proof exists to support the unbiasedness of the B-W scores, empirical study has demonstrate that "this is of Multinomial Logit nature in terms of ratios of scale values and the scores are a sufficient statistic for parameter estimation" (*8, 9*).

Given the design used in our survey, where all combinations were presented evenly, we can go further and use the B-W scores on an individual level, with an assumption that the attribute-specific B-W scores for each respondent also provides us with an indication of how this particular respondent attaches importance to each attribute. As each attribute appeared 4 times for each respondent, the individual-level B-W score for any attribute could range from -4 to 4.

Results

Table 1 summarises the aggregated B-W score for each attribute across respondents at sample level as well as the standard deviation (s.d.) of individual-level B-W scores for each attribute. On the one hand, the B-W scores across respondents for each attribute provide a straightforward implication that minor time and ticket integration are considered as the least important while connection time and travel cost are the two most important attributes at sample level when respondents need to decide whether to buy an integrated HSR-air travel service. On the other hand, the standard deviations of B-W scores suggest that respondents' importance of the last 4 attributes are more diverse than those on the three time components. Minor time has the lowest B-W scores and is the attribute with the second lowest standard deviation of B-W scores in the sample, which indicates that minor time is universally considered as unimportant. This is understandable as our survey was based in Shanghai and its nearby regions which could be reached by HSR or air from Shanghai within a relatively short time, therefore respondents may feel that minor time is not important. A somewhat different picture emerges later on in the choice models, with relatively high time sensitivities, a point we will return to below.

TABLE 1 Aggregated B-W scores and Standard Deviation of B-W Scores in the sample

#	Attribute	B-W score	s.d.	Ranking
1	minor time	-111	1.77	7
2	connection time	45	1.99	1
3	transfer time	28	1.76	4
4	delay protection	36	2.34	3
5	ticket integration	-58	2.26	6
6	luggage integration	20	2.60	5
7	travel cost	40	2.49	2

We next conduct a comparative analysis between the individual-specific B-W scores and the observed choice outcomes. We look at the frequency of choosing the alternative with the lowest minor time, the one with the lowest connection time, the one with the lowest transfer time, and so forth. We do not find very strong correlation between the B-W scores and these choice strategies, but the weak links between the two can still provide us with some useful indications about attribute importance (the correlation coefficients mentioned below are all significant at 95% confidence level). It should be noted here that obviously more than one can apply at the same time in one choice (e.g. the fastest may also have the shortest connection time). We see that the B-W score on delay protection is positively correlated with the frequency of choosing the highest delay protection ($\rho = 0.29$); the B-W score on luggage integration is positively correlated with the frequency of choosing the highest luggage integration ($\rho = 0.17$), and negatively correlated with the frequency of choosing the lowest travel cost ($\rho = -0.25$). This means that those who have higher B-W scores on delay protection are more frequently observed to choose the alternative with the highest level of delay protection; and respondents with higher B-W scores on luggage integration choose the alternative that can provide best integration service more often, and meanwhile care less about travel cost.

We also compare the individual-specific B-W scores against the frequency of each alternative being chosen in the SC survey. Again, only weak but significant correlation is detected where some useful implication can still be extracted. Firstly, it is discovered that the B-W score on connection time is positively correlated with the frequency of the separated HSR-air alternative being chosen in the SC tasks ($\rho = 0.33$), and negatively correlated to the choice frequency for any of the other three alternatives (car-air: $\rho = -0.16$; air-air: $\rho = -0.18$; integrated HSR-air: $\rho = -0.19$). Second, higher counts on luggage integration is related to lower frequency of choosing separated HSR-air ($\rho = -0.32$) and higher frequency of choosing integrated HSR-air ($\rho = 0.22$). These two relationships might result from the fact that the separated HSR-air travel could provide more flexibility to passengers by allowing them to have more control over the travel themselves and shorten the waiting time between the major and minor leg, whereas the integrated counterpart might "force" those passengers to spend more time on waiting and use the integrated luggage handling service which is not required.

ANALYSIS II: POSTERIORS FROM BAYESIAN ESTIMATION

Method

Our next analysis obtains individual-specific posteriors from a Mixed Multinomial Logit (MMNL) analysis of the stated choice data and contrasts these with the B-W scores. Following the procedures proposed in (*11, 12*), we use Bayesian estimation of a MMNL model where we allow for random variation in all parameters, with correlation between individual parameters.

In the model specification, the utility that respondent n obtains from alternative i at choice task t is given as $U_{int} = ASC_{in} + \beta_n' x_{int} + \varepsilon_{int}$, with β_n being the vector of taste coefficients for respondent n and ε_{int} being iid extreme value. We constrain the coefficients for the alternative attributes to take the expected sign for all respondents by assuming positive Log-normal distribution for "good attributes" including delay protection, ticket integration and luggage integration $(k = 4, 5, 6)$, such that:

$$\beta_{nk} = e^{\mu_{ln(\beta_{nk})} + \sigma_{nk}\xi_k} \tag{1}$$

and negative Log-normal distribution for "bad attributes" including minor time, connection time, transfer time and travel cost $(k = 1, 2, 3, 7)$, in a form of:

$$\beta_{nk} = -e^{\mu_{ln(-\beta_{nk})} + \sigma_{nk}\xi_k} \tag{2}$$

where μ and σ are the to-be-estimated means and standard deviations for the underlying Normal distribution. ξ_k follows a standard Normal distribution across respondents for attribute k, such that $\xi_k \sim N(0, 1)$.

The three alternative-specific constants (ASC) are specified to follow Normal distribution, to account for the underlying preference of the specific alternative which might be above or below the base alternative (i.e. integrated HSR-air[2]) given all else being equal. Minor time is separated between car or air and HSR; besides, different levels of some attributes, including delay protection,

[2]According to (*13*), the integrated HSR-air is chosen as the base alternative as it has the lowest variance in an unidentified model.

ticket integration, and luggage integration, are dummy coded with constraints that the utility sensitivity is monotonous for each attribute across the levels by using additive Log-normal distributions to assure that higher level is better than the lower level for these attributes.

The models are estimated by using a panel formulation which assumes that sensitivities vary across respondents but stay constant across choice tasks for each respondent. The Bayesian estimation is conducted in RSGHB (*14*), with 2,000,000 iterations in the burn-in procedure to use prior to convergence and another 200,000 iterations for averaging after convergence has been reached and we retain every fifth draw for averaging.

Let $P_{nt}(i_{nt} \mid \beta_n)$ denote the conditional probability of respondent n choosing alternative i at choice task t given a specific value of β_n, which has a prior Normal density $f(\beta_n \mid \theta)$ with θ representing the collective of distributional parameters. We label the sequence of choices for respondent n as y_n and then the probability of observing y_n given β_n is denoted as $P_n(y_n \mid \beta_n)$. The marginal probability of observing y_n is given as the integral of the probability of the choice sequence conditional on β_n over the prior distribution of β_n, such that:

$$\begin{aligned} P_n(y_n) &= \int_{\beta_n} P_n(y_n \mid \beta_n) f(\beta_n \mid \theta) d\beta_n \\ &= \int_{\beta_n} \prod_{t=1}^{T} P_{nt}(i_{nt} \mid \beta_n) f(\beta_n \mid \theta) d\beta_n \end{aligned} \tag{3}$$

Based on Bayes' rule, we can have the possibility of observing a specific value of β_n for respondent n given the observed choices y_n is:

$$P_n(\beta_n \mid y_n) = \frac{P_n(y_n \mid \beta_n) f(\beta_n \mid \theta)}{P_n(y_n)} \tag{4}$$

which is also called posterior distribution. The mean of the posterior distribution for person n, which reflects the most likely value for the parameters given the observed choices for this person, is then given as:

$$\hat{\beta}_n = \frac{\sum_{r=1}^{R}[P(y_n \mid \beta_r)\beta_r]}{\sum_{r=1}^{R} P(y_n \mid \beta_r)} \tag{5}$$

where β_r with $r = 1, ..., R$ are independent multi-dimensional draws with equal weight from $f(\beta \mid \theta)$ at the estimated values for θ (*15*).

Results

Since posterior distributions are inferred from the SC data itself and B-W scores are information obtained from respondents' self assessment, we can thereby bridge the understandings of attribute importance from these two different sides and also compare the inferred results with observed choice outcomes. We make use of the individual-specific mean of the posterior distribution for each attribute and analyse its correlation with the individual-specific B-W scores for each attribute, as shown in Figure 2, with the number in each cell giving the Pearson correlation between the corresponding row and column, where blue cells stand for positive correlations and red cells for negative correlations. For "good attributes", the figure suggests positive correlations with the means of posterior distributions for almost all the sensitivity coefficients, in that higher B-W scores can be linked with more positive sensitivities of "good attributes", and the converse applies for "bad attributes". This means for example that if a respondent is observed to have higher B-W score on luggage integration, the mean of the posterior distribution for this coefficient is likely to be higher. We also see that there is positive correlation across the "good attributes", indicating that someone who attaches high importance to some qualitative attributes is likely to do the same for others. The same rationale applies for "bad attributes". This finding is also in accordance with our intuitive expectation that passengers who attach more importance to travel time or travel cost would be more restricted by the duration or the expenditure of the travel and meanwhile derive less positive utilities from those "good attributes". For instance, those observed to have higher B-W scores on connection time are inferred to be more affected by the constraints on connection time or transfer time, and derive less positive utilities from the extra services provided by "good attributes".

The presence of some weaker correlations between B-W scores and inferred sensitivity coefficients in Figure 2, like the results in our first analysis, suggests a probability of some inconsistency between passengers' responses to B-W tasks and SC tasks for a subset of the attribute package, which might be the result of respondents rating attributes differently when not faced with a multi-alternative trade-off where they have to accept bad performance for some attributes in return for good performance for other.

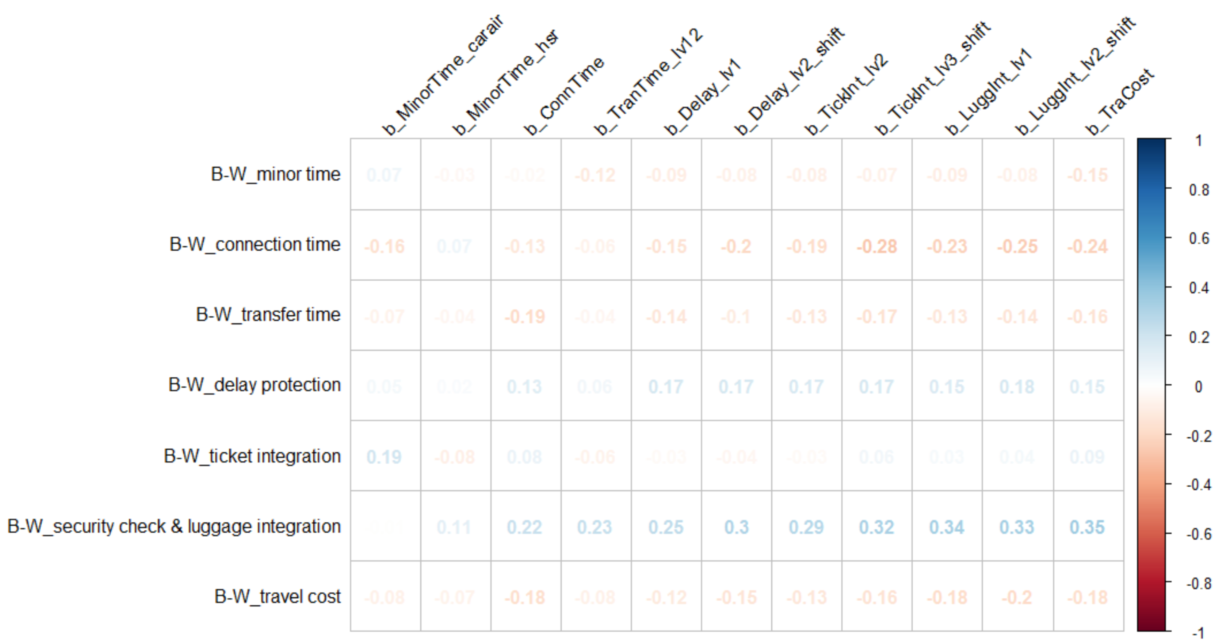

FIGURE 2 Correlation between BW scores and posterior sensitivities.

ANALYSIS III: HYBRID CHOICE MODEL APPROACH

Method

We finally make use of a hybrid choice model based on the concept of latent attribute importance, which jointly explains taste heterogeneity in the choice model and the values of the B-W scores. This is analogous to the approach adopted in (*1*) and builds on the general hybrid framework of (*16*). Figure 3 provides an illustration of our model structure, where utilities are determined by both observable characteristics of alternatives and latent variables of attribute importance. The model consists of two parts, which are a choice model component and a latent variable component, each including structural equations and measurement equations. Items in rectangular are observable to researchers and items in ellipse are unobserved. Solid arrows represent structural equations which describe the causal relationship between unobserved items and observed items, while dashed arrows refer to measurement equations which explain indicators by latent variables or choices by utilities.

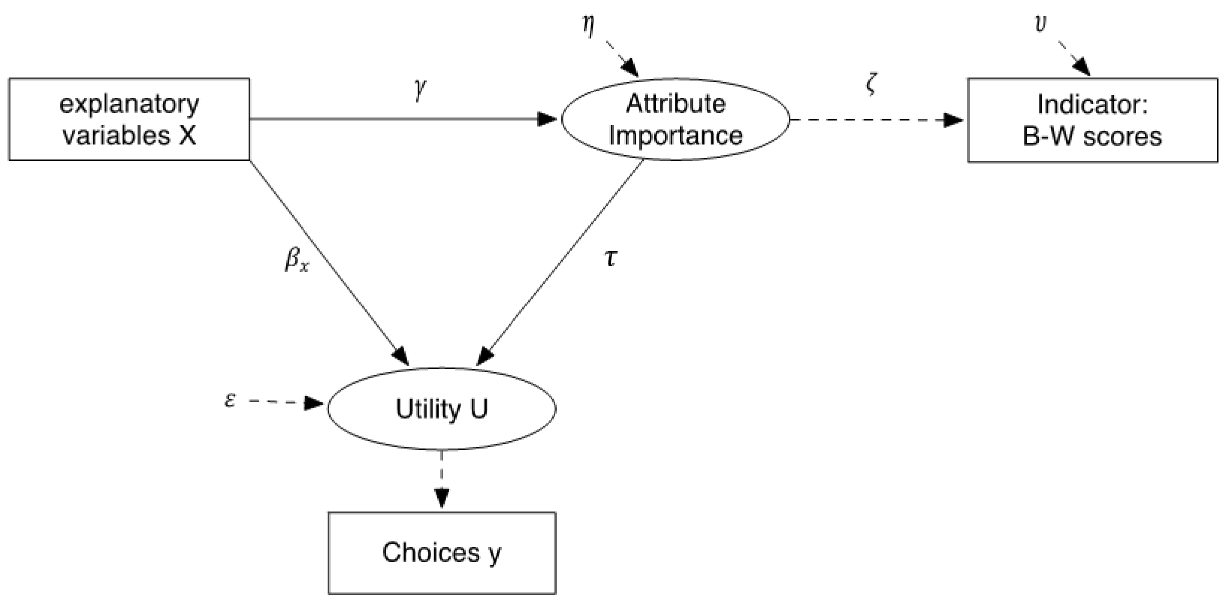

FIGURE 3 Framework of the HCM model.

Since seven attributes are included in our survey, seven latent variables, each corresponding to a particular attribute, are defined here which are: α_1 for minor time, α_2 for connection time, α_3 for transfer time, α_4 for delay protection, α_5 for ticket integration, α_6 for luggage integration and α_7 for travel cost. The latent attribute importance is used to explain both the sensitivities to individual attributes in the utility function and the responses to indicators in the measurement equations, where the corresponding individual-specific B-W score is used as the indicator. In this exploratory work, we do not incorporate a deterministic component into the structural equation, and thus assume pure randomness of the latent variable across respondents and specify Normal distribution for each latent variable, such that:

$$\alpha_{nk} = \eta_{nk} \tag{6}$$

where $\eta_{nk} \sim N(0,1)$.

We adopt a random coefficients formulation which allows for heterogeneous preference coefficients in addition to the impacts of latent variables across respondents, while maintaining homogeneity within a respondent across all choice tasks. Similar to Analysis II, Log-normal distributions are specified for all the attribute coefficients β_{nk} to assure the expected signs being taken by all respondents. Monotonic constraints are applied to the different levels of delay protection, ticket integration and luggage integration, where we allow for different means for the underlying

Normals but due to limited data rely on the same variance of the underlying Normal distribution for different levels of k. Correlations are not specified between different underlying Normal distribution for the same reason. We specify an exponential multiplier for attribute importance, such that the separate random utility coefficient for attribute k is multiplied by $e^{\tau_k \alpha_{nk}}$, where τ_k measures the impact of latent α_{nk} on scaling the sensitivity coefficients β_{nk} inside the choice model. As the latent variable has a Normal error term (see Equation 6), the scaled sensitivity coefficients still follow a Log-normal distribution.

We then have:

$$\begin{cases} \beta_{n1,ca} = -e^{\tau_1 \alpha_1} e^{\mu_{ln(-\beta_{n1,ca})} + \sigma_1 \xi_1} \\ \beta_{n1,h} = -e^{\tau_1 \alpha_1} e^{\mu_{ln(-\beta_{n1,h})} + \sigma_1 \xi_1} \\ \beta_{n2} = -e^{\tau_2 \alpha_2} e^{\mu_{ln(-\beta_{n2})} + \sigma_2 \xi_2} \\ \beta_{n3} = -e^{\tau_3 \alpha_3} e^{\mu_{ln(-\beta_{n3})} + \sigma_3 \xi_3} \\ \beta_{n4,1} = e^{\tau_4 \alpha_4} e^{\mu_{ln(\beta_{n4,1})} + \sigma_4 \xi_4} \\ \beta_{n4,2\_shift} = \beta_{n4,1} + e^{\tau_4 \alpha_4} e^{\mu_{ln(\beta_{n4,2\_shift})} + \sigma_4 \xi_4} \\ \beta_{n5,2} = e^{\tau_5 \alpha_5} e^{\mu_{ln(\beta_{n5,2})} + \sigma_5 \xi_5} \\ \beta_{n5,3\_shift} = \beta_{n5,2} + e^{\tau_5 \alpha_5} e^{\mu_{ln(\beta_{n5,3\_shift})} + \sigma_5 \xi_5} \\ \beta_{n6,1} = e^{\tau_6 \alpha_6} e^{\mu_{ln(\beta_{n6,1})} + \sigma_6 \xi_6} \\ \beta_{n6,2\_shift} = \beta_{n6,1} + e^{\tau_6 \alpha_6} e^{\mu_{ln(\beta_{n6,2\_shift})} + \sigma_6 \xi_6} \\ \beta_{n7} = -e^{\tau_7 \alpha_7} e^{\mu_{ln(-\beta_{n7})} + \sigma_7 \xi_7} \end{cases} \tag{7}$$

where the subscript k after n stands for the attribute (i.e, minor time: $k = 1$, connection time: $k = 2$, transfer time: $k = 3$, delay protection: $k = 4$, ticket integration: $k = 5$, luggage integration: $k = 6$, travel cost: $k = 7$). The subscript after the comma in β_{n4}, β_{n5}, and β_{n6} relates to different levels of the attribute, while in β_{n1}, it stands for the separate estimates for car or air and for HSR.

In the measurement equations, the individual-specific B-W scores I_{nk} are treated as indicators of the corresponding latent variable α_k and each indicator requires a separate measurement equation. Although ordered Logit specifications in measurement equation (*17*) have been advocated in recent years to account for the ordered nature of responses to attitudinal statements, we still adopt the traditional linear specification as our B-W scores are not responses on a Likert scale and may range from -4 to 4, such that a large number of parameters would need to be estimated with sparse data. The measurement equations can thus be modelled as:

$$I_{nk} = \zeta_k \alpha_{nk} + \upsilon_{nk} \tag{8}$$

where ζ are the to-be-estimated parameters that reflect the impacts of latent variables on B-W score indicators. The random term υ_{nk} is assumed to follow a Normal distribution with a mean of zero, such that $\upsilon_{nk} \sim N(0, \varsigma)$ with ς being the standard deviation to be estimated.

Log-likelihood maximisation is adopted for estimation, such that $max(LL(Y, I))$, where we need to maximise the log-likelihood of observing the choices Y and indicators I. The unconditional probability of observing choices Y and indicators I can be expressed as the integral of the multiplication of conditional choice probability and the conditional indicator probability over the distribution of the latent variables, such that:

$$LL(Y,I) = \sum_{n=1}^{N} ln \int_{\beta_n} \int_{\alpha_n} (\prod_{t=1}^{T_n} P(y_{nt} \mid x_{nt}, a_n) \times \prod_{k=1}^{K_n} P(I_{nk} \mid \alpha_n)) f(\alpha_n) d\alpha_n f(\beta_n \mid \theta) d\beta_n \quad (9)$$

As random coefficients are accounted for within a panel formulation, a second layer of integral over all possible values of β is required. Since the resulting LL does not have closed-form expression, the estimation needs to be approximated through simulation (*11*). The presence of the separate layer of random heterogeneity ensures that we do not misattribute heterogeneity to the latent variables but are able to disentangle a random part which is linked to the latent variable and a part which is not.

Estimation results

The estimation results of the hybrid choice model are presented in Table 2, where items in bold are significant at 95% confidence level. The significant estimates of the three alternative specific constants suggest the existence of underlying preference for these alternatives, where we do not in the present work allow for additional heterogeneity in these constants.

We first look at the estimates for the measurement equations before turning to the impact of the latent variables on scaling utility sensitivities in the choice model component. It is shown that ζ_4, ζ_5, ζ_6, and ζ_7 are significant at 95% confidence level and ζ_2 is significant at 85% level, which suggests that the indicators of B-W scores for attributes of delay protection, ticket integration, luggage integration and travel cost and potentially connection time are significantly affected by the corresponding latent variables. The positive signs for ζ_2, ζ_4, ζ_6, and ζ_7 and negative sign for ζ_5 show that stronger latent α_2, α_4, α_6, α_7 and weaker α_5 would lead to an increase in the corresponding B-W score. This also suggests that α_2, α_4, α_6, α_7 actually stands for "attribute importance" of connection time, delay protection, luggage integration and travel cost respectively, while α_5 for "attribute unimportance" of ticket integration. On the contrary, the impacts for latent variables α_1 and α_3 on the corresponding B-W score indicators are not clear (ζ_1: t-stat=-0.11, ζ_3: t-stat=0.49). Since minor time has the lowest aggregated B-W counts and transfer time has the lowest standard deviation of B-W scores (see Table 1), it may suggest that the majority of respondents view minor time as very unimportant in decision making and have the least difference in the opinions on transfer time, which could potentially result in the insignificant impacts of latent attribute importance on the B-W scores.

Turning to the impacts of latent variables in the choice model, it is shown that τ are significantly estimated for all the attributes except for ticket integration (τ_5: t-stat=-0.43), revealing the presence of scaling effect introduced by latent variables on attribute importance, which confirms the findings in previous research (*1*). The negative sign for minor time (τ_1) and the positive signs for the remains imply that a decrease in latent variable α_1 and increases in the latent variable α_2, α_3, α_4, α_6 and α_7 can lead to stronger utility sensitivities for the attribute concerned. Such results are generally in accordance with our expectations, as earlier interpretation of α_2, α_4, α_6 and α_7 as "attribute importance" shows that stronger attribute importance attached to connection time, delay protection, luggage integration and travel cost leads to stronger scaling effect and thus higher marginal utilities on concerned attribute, while weaker attribute importance results in a higher possibility that the concerned attribute is ignored or ranked as less important. In addition, though the corresponding impacts of latent variables on B-W indicators are not significantly estimated in respect of minor time and transfer time (see ζ_1 and ζ_3), the significant τ_1 and τ_3 together with the significant corresponding variances ς_1 and ς_3 still manifest the presence of scaling effect for the attributes of minor time and transfer time, which is purely random and irrelevant to the latent variable, making it difficult to define what latent constructs α_1 and α_3 actually stand for.

Turning to the estimates of the underlying Normal distributions for the utility sensitivity coefficients, all the underlying means except for $\mu_{ln(-\beta_{5\_3\_shift})}$ and all the underlying variances except for σ_5 are significant at 90% level at least, suggesting the presence of random heterogeneity independent of the latent variables. In addition to the random heterogeneity in the β parameters, we also see an impact by the latent variable through the τ parameter. These need to be interpreted alongside the ζ parameters. We can observe that for delay protection, luggage integration and travel cost, increases in the latent variable lead to higher B-W scores as well as increases in the absolute value of β, supporting a link between attribute importance in the SC data and the B-W scores. A weaker link exists for connection time, where the ζ term is only marginally significant but τ is highly significant.

CONCLUSIONS

This paper has sought to make a link to respondents answers on attribute importance using B-W scaling and their behaviour in a stated choice survey. This builds on earlier work creating a link between stated attribute importance and taste heterogeneity, but using a different and arguably richer response mechanism in the form of B-W scaling, which is growing in popularity across different fields. At the outset, it should already be acknowledged that a key issue in this context arises for continuous attributes as well as multi-level categorical attributes. A respondent may rank an attribute as important, but the impact on choices will depend on the specific values taken by the attribute.

Our analysis uses three distinct approaches to look for links, starting with a descriptive analysis, followed by an investigation using posterior distributions from a Mixed Logit model and culminating in the use of a hybrid choice model. The exploratory comparison analysis suggests weak correlation between B-W scores and stated choice outcomes, where in particular respondents with higher B-W score on delay protection and luggage integration are more frequently observed to select the alternative that provides the highest level of delay protection and the best luggage integration in respective. We also show that respondents who choose the separated HSR-air more often have a stronger dislike for longer connection time, and are less attracted by luggage integration, which is potentially due to that they want to have more flexible travel and shorten the waiting time at the airport for the major leg flight. The correlation analysis between the Bayesian posteriors and B-W scores, both varying across respondents and across attributes, indicate positive correlation between the B-W scores on some "good attributes" and the means of posterior distributions for either good or bad attributes, which means that respondents with higher B-W scores on "good attributes" (including delay protection and luggage integration) are less restricted to "bad attributes" (including connection time, transfer time and travel cost) and are more willing to pay for the extra service offered by these "good attributes" at either monetary or temporal cost. Finally, our hybrid choice model shows that for four attributes (including connection time, delay protection, luggage integration and travel cost), there is an impact by the latent attribute importance on both the parameter magnitude and the associated B-W score.

There are some avenues for future research. Firstly, it is necessary to explore the possible causal relationships between the latent variables and observable explanatory variables, such as

TABLE 2 Estimation Results of Choice Model

$LL(overall)$	-2897.1490				
$LL(choice)$	-1033.7510				
$Parameter\#$	42				
$Choice\ model$	$Est.$	$robtrat\_0$	$Measurement\ equation$	$Est.$	$robtrat\_0$
ASC_1	-1.9184	**-3.32**	ζ_1	-0.0497	-0.11
ASC_2	0.9605	**4.37**	ζ_2	0.3881	1.46
ASC_3	-0.9467	**-4.26**	ζ_3	0.1824	0.49
$\mu_{ln(-\beta_{n1\_ca})}$	-4.4591	**-14.28**	ζ_4	1.2100	**2.38**
$\mu_{ln(-\beta_{n1\_h})}$	-6.2729	**-12.71**	ζ_5	-1.2910	**-3.63**
$\mu_{ln(-\beta_{n2})}$	-4.4440	**-35.22**	ζ_6	1.5603	**4.38**
$\mu_{ln(-\beta_{n3})}$	-0.4295	-1.68	ζ_7	0.7466	**2.09**
$\mu_{ln(\beta_{n4\_1})}$	-1.8489	-1.91	ς_1	1.7727	**17.36**
$\mu_{ln(\beta_{n4\_2\_shift})}$	-2.6082	-1.65	ς_2	1.9553	**17.34**
$\mu_{ln(\beta_{n5\_2})}$	-10.5726	-1.69	ς_3	1.7513	**14.39**
$\mu_{ln(\beta_{n5\_3\_shift})}$	-10.6946	-1.07	ς_4	1.9880	**6.16**
$\mu_{ln(\beta_{n6\_1})}$	-0.8108	**-2.33**	ς_5	1.8827	**8.91**
$\mu_{ln(\beta_{n6\_2\_shift})}$	-4.6191	-1.49	ς_6	2.0626	**7.99**
$\mu_{ln(-\beta_{n7})}$	-6.0995	**-28.61**	ς_7	2.3746	**16.81**
σ_1	-1.0760	**-5.23**			
σ_2	0.2381	**3.95**			
σ_3	0.4422	**2.34**			
σ_4	-0.4678	**-3.43**			
σ_5	-0.9768	-0.54			
σ_6	-0.7076	**-5.25**			
σ_7	0.4660	1.74			
τ_1	-0.6332	**-5.02**			
τ_2	0.5496	**6.76**			
τ_3	0.6741	**6.27**			
τ_4	1.0866	**4.59**			
τ_5	-1.8921	-0.43			
τ_6	1.2029	**5.45**			
τ_7	0.5023	**4.73**			

different socioeconomic characteristics, to improve the structural equations. Secondly, we are exploring other specifications of incorporating the latent attribute importance in choice model at the moment, and this would help us to better understand the impact of attribute importance. Thirdly, as we also collected passengers' responses to another set of BWS case 2 tasks in the survey along with the SC tasks and BWS case 1 tasks, where the to-be-evaluated items are a list of attribute levels for different attributes, it would be possible to further analyse the importance of attribute levels with an assumption that for some respondents, certain attribute levels are ignored or considered as less important for choice decisions. Finally, it would be of interest to investigate why some inconsistency exists between how respondents perceive attribute importance in BW tasks and in traditional SC tasks.

ACKNOWLEDGMENT

Fangqing Song acknowledges the support of the China Scholarship Council while Stephane Hess was supported by the European Research Council through the consolidator grant 615596-DECISIONS.

REFERENCES

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

2. Terry N Flynn, Jordan J Louviere, Tim J Peters, and Joanna Coast. Estimating preferences for a dermatology consultation using best-worst scaling: comparison of various methods of analysis. BMC Medical Research Methodology, 8(1):76, 2008.

3. Dimitris Potoglou, Peter Burge, Terry Flynn, Ann Netten, Juliette Malley, Julien Forder, and John E Brazier. Best–worst scaling vs. discrete choice experiments: an empirical comparison using social care data. Social science & medicine, 72(10):1717–1727, 2011.

4. Jennifer A Whitty, Julie Ratcliffe, Gang Chen, and Paul A Scuffham. Australian public preferences for the funding of new health technologies: a comparison of discrete choice and profile case best-worst scaling methods. Medical Decision Making, 34(5):638–654, 2014.

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

6. John M Rose and Michiel CJ Bliemer. Stated preference experimental design strategies. In Handbook of Transport Modelling: 2nd Edition, pages 151–180. Emerald Group Publishing Limited, 2007.

7. Choice Metrics. Ngene 1.1. 1 user manual & reference guide. Sydney, Australia: ChoiceMetrics, 2012.

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

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

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

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

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

13. Joan L Walker, Moshe Ben-Akiva, and Denis Bolduc. Identification of parameters in normal error component logit-mixture (neclm) models. Journal of Applied Econometrics, 22(6):1095–1125, 2007.

14. Jeff Dumont, Jeff Keller, Chase Carpenter, and Maintainer Jeff Dumont. Package rsghb. 2015.

15. Stephane Hess. Conditional parameter estimates from mixed logit models: distributional assumptions and a free software tool. Journal of Choice Modelling, 3(2):134–152, 2010.

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

17. Andrew Daly, Stephane Hess, Bhanu Patruni, Dimitris Potoglou, and Charlene Rohr. Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour. Transportation, 39(2):267–297, 2012.