Discussion Paper Series

RIEB

Kobe University

DP2019-14

An Empirical Comparison Between Discrete Choice Experiment and Best-worst Scaling: A Case Study of Mobile Payment Choice

Qinxin GUO
Junyi SHEN

July 10, 2019

Research Institute for Economics and Business Administration
Kobe University
2-1 Rokkodai, Nada, Kobe 657-8501 JAPAN

An empirical comparison between discrete choice experiment and best-worst scaling: A case study of mobile payment choice

Qinxin Guo[a,*], **Junyi Shen**[b,c]

[a] Graduate School of Economics, Kobe University, Kobe, 2-1 Rokkodai, Nada, Kobe 657-8501, Japan.

[b] Research Institute for Economics and Business Administration, Kobe University, 2-1 Rokkodai, Nada, Kobe 657-8501, Japan.

[c] School of Economics, Shanghai University, 99 Shangda Road, Baoshan 200444, Shanghai, China

[*] Corresponding author. Email: guoguoxionggqx@gmail.com

Abstract

As an alternative method to discrete choice experiments, best-worst scaling provides additional information about consumers, slightly lessens the burden of mental process, and shows better quality. However, its advantages were ambiguous in previous literature, since each case of the best-worst scaling contained distinct information, and results from comparisons with discrete choice experiment varied with different data. In this study, we applied a goodness of fit statistic named count R square in evaluating the best-worst scaling profile case, the discrete choice experiment, and the best-worst scaling multi-profile case by using data from a survey of preference for mobile payment. The results suggest that the best-worst multi-profile case surpasses other methods. We also compared the mixed logit model and the latent class model using three non-nested tests. The results indicate that the mixed logit model is superior to the latent class model in all three tests.

Keywords: Discrete choice experiment; Best-worst scaling; Goodness of fit; Latent class model; Mixed logit model

1. Introduction

Since the discrete choice experiment (DCE) was developed in the early 1980s (Louviere and Woodworth, 1983), it has been a workhorse to evaluate the stated preferences for various issues in the economics literature (in the areas of transportation economics, health economics, environmental economics, etc.). The DCE was implemented within the random utility theory (Thurstone, 1927), which depicted the choice behavior from the perspective of economic theory relatively well. Hence, the DCE method surpassed other stated preference methods that were not derived from the economic theory and became a popular analysis tool in the economics literature. However, due to the absence of a ranking system of alternatives, DCE barely allowed us to know the "best choice" for respondents. Thus, scholars sought to further investigate and acquire more information about the choice behavior.

There has been an increasing interest in an alternative method called the best-worst scaling (BWS) to elicit more information based on DCE in recent decades. Finn and Louviere (1992) first proposed the BWS method for a food safety case, in which a person was asked to select both the best and worst items from a list of options in terms of food safety. Since the pilot research was published, a number of applications have been proposed and a complete theoretical system has been established over the last decades. The BWS method basket includes three types, object case, profile case, and multi-profile case (Flynn, 2010). In the object case, respondents choose the best and worst objects from a list of objects (or attributes without detailed levels). The profile case involves only one profile or alternative in a normal DCE choice set and respondents choose the best and worst levels from this profile. The multi-profile case adds the question of which profile the respondents like least to each choice set of DCE, in addition to the question of which profile the respondents like the most. Namely, the respondents need to choose the best and worst profiles from each choice set in the BWS multi-profile case.

The BWS method seemed attractive because it provides more information and lowers the cognitive burden (e.g., object case and profile case) compared with the traditional DCE method (Potoglou et al., 2011). However, results of previous studies indicated that the superiority of BWS was questionable when utilizing different data. Moreover, economists primarily focused on the comparison between DCE and two types of BWS (i.e., BWS profile case and BWS multi-profile case). The BWS profile case stood out from the aforementioned three methods, since the form of the BWS profile case allow respondents to make a choice regarding one profile instead of selecting between two or more profiles as in the other two methods (Louviere et al., 2015). The easier choice set of the BWS profile case might allow respondents to make their choice more quickly. Meanwhile, the BWS profile case separates a single profile, which allows the respondents to compare the attribute levels more explicitly. Nevertheless, statistical results did not provide a clear proof that one of the aforementioned methods is better than the others, since the results vary with different datasets and models. Furthermore, some previous studies compared the BWS and DCE using choice certainty (Yoo and Doiron, 2013; Flynn et al., 2013; Whitty et al., 2014; Xie et al., 2014). The authors of these studies found that parameters of logit regression contain information on the decision certainty of respondents, while the higher scale parameters imply lower variance of error terms, which suggests respondents express more decision certainty when making decisions[1]. Nonetheless, the comparison results are different for various datasets and econometric models. In this study, we compare the aforementioned three methods using an empirical case (i.e., the mobile payment choice) and utilized a goodness of fit measure named count R square (i.e., the ratio of the number of correct predicted outcomes to the number of total predicted outcomes in the regression model) as the criteria to determine which method has a better explanatory power for respondents' stated preference. Our results indicate that the more information included in the data as in the BWS multi-profile case, the closer the regression result is to respondents' real choice behavior. In

[1] In the next section, we can see that both DCE and BWS are based on random utility theory, i.e. $U_{iq} = V_{iq} + \varepsilon_{iq}$. The error term ε_{iq} is usually assumed to follow Gumbel distribution: $F(e_{iq}) = \exp(-\exp(-\mu\varepsilon_{iq})),\ \mu > 0$. Therefore, the variance of error term is presented as: $V(e_{iq}) = \pi^2 / 6\mu^2$, where μ is the scale parameter (Ben-Akiva and Lerman, 1985). Although the choice of μ is arbitrary, since it simply sets the scale of the utilities, the fact that each error term has the same value of μ implies the variances of the random components of the utilities are equal, and that the larger scaled parameters have lower uncertainty according to comparisons in past literature.

addition, we also conduct a comparison between the Mixed Logit Model (MLM) and Latent Class Model (LCM) specifications using all three methods. The results suggest that the MLM specification is superior to the LCM one.

The remainder of this paper is organized as follows: The next section describes the methodology of DCE and BWS, and survey design and data collection are presented in Section 3. Section 4 offers results of the comparison between the two BWS cases and DCE. Section 5 provides results of the comparison between MLM and LCM specifications with conclusions presented in the final section.

2. Methodology

Both discrete choice experiment and best-worst scaling are based on the random utility theory; hence, the models for the two methods are similar except for the additional information that is included in the BWS method. The basic assumption of random utility theory is that decision makers maximize their utility by choosing their favorite alternative among a set of alternatives (Shen, 2006). The real utility of an alternative for an individual U cannot be observed; however, it could be seen as consisting of a deterministic component V and a random error term ε. An individual q choosing an alternative i can be expressed as:

$$U_{iq} = V_{iq} + \varepsilon_{iq} \tag{1}$$

The probability of individual q choosing an alternative i from a set J that contains j alternatives can be written as:

$$P_{iq} = P(U_{iq} > U_{jq}; \forall i \neq j \in J) = P(\varepsilon_{jq} < \varepsilon_{iq} + V_{iq} - V_{jq}; \forall i \neq j \in J) \tag{2}$$

Additionally, both discrete choice experiment and best-worst scaling interpret the aforementioned choice process with the error terms following Gumbel distribution and obtain a Multinomial (or conditional) logit model. Therefore, we could rewrite Equation (2) as:

$$P_{iq} = \exp(\mu V_{iq}) / \sum_{j=1}^{J} \exp(\mu V_{jq}) \tag{3}$$

2.1 Discrete choice experiment

In the DCE, decision makers simply choose their favorite alternative among a set of alternatives in the choice sets, while the observable utility value could be presented as linear in parameters, $V_{iq} = \beta' X_{iq}$. Thus, Equation (3) could be shown as:

$$P_{iq} = \exp(\mu\beta' X_{iq}) / \sum_{j=1}^{J} \exp(\mu\beta' X_{jq}) \qquad (4)$$

where μ represents a parameter that determines the scale of the utilities, which is proportional to the inverse of the distribution of the error terms. X_{iq} are explanatory variables of V_{iq} and usually include alternative-specific constants (ASCs), the attributes of the alternative i, and the social characteristics of individual q. β' are parameter vectors associated with X_{iq}. Regarding the logit model of DCE method, we can only capture information about the favorite alternative, but might lose additional crucial information about alternatives that decision makers do not favor, since the choice is only made by selecting the favorite alternative. Conversely, the BWS method allows decision makers to choose another option that they like the least, while the logit model in this method mainly applies the difference between the best choice and the worst choice to interpret the decision process.

2.2 Best-worst scaling

Case 1 (object case)

The object case was proposed by Finn and Louviere (1992) as the pilot of BWS method. Object case requires a “list” of objects (or attributes without levels) that one wants to measure, and the decision makers need to choose both the best and worst objects in the choice set (Louviere et al., 2015).

Since the theoretical framework of the BWS method is similar to that of DCE, the random utility theory is presented through the maxdiff model in the BWS method. In the maxdiff model, the potential best and worst choices are defined as a pair, and the error term is assumed to follow the Gumbel distribution for every pair of this best-worst choice combination. X is the choice set, while m are the attributes ($m \geq 2$). Let $M = \{1, \ldots, m\}$ and assume attribute i, $i = 1, \ldots, m$. Then, a maxdiff model of best-worst choice probabilities in object case is given by the following equation:

$$P_{BW}(ii' \mid X) = \exp(\mu(V(i) - V(i'))) / \sum_{\substack{j, j' \in M \\ j' \neq j}} \exp(\mu(V(j) - V(j'))) \qquad (5)$$

where i stands for the one that decision makers choose as the best item, while i' stands for the one that decision makers choose as the worst item. μ represents a parameter that determines the scale of the utilities and $V(i)$ is the utility of item i (or attribute i). Moreover, the maxdiff model assumes that decision makers simultaneously choose the best-worst items as a pair; however, when individuals make decisions, it might be sequential (i.e., first best then worst or first worst then best). Therefore, in the profile case and multi-profile case, it is usually beneficial to consider the sequential form in the analysis.

Case 2 (profile case)

The profile case choice set contains a single profile (i.e., alternative) and the decision makers need to choose the best attribute level and the attribute level and the worst attribute level in that specific profile. Regarding previous notation, attribute i is assumed to have $q(i)$ levels in profile case, while the profile is an m-component vector with each component i taking on one of the $q(i)$ levels for that component. A profile j is denoted by x_j, $x_j = (x_{j1}, \ldots, x_{jm})$; thus, the maxdiff model of best-worst choice probabilities in profile case can be presented as:

$$P_{BW}(ii' \mid x_k) = \exp(\mu(\beta_i x_{ki} - \beta_{i'} x_{ki'})) / \sum_{\substack{j, j' \in M \\ j' \neq j}} \exp(\mu(\beta_j x_{kj} - \beta_{j'} x_{kj'})) \qquad (6)$$

where x_k is profile k and x_{ki} is the attribute level chosen as the best option and $x_{ki'}$, while $i \neq i'$ is the attribute level chosen as the worst option in profile k. μ represents a parameter that determines the scale of the utilities, while β_i are parameter vectors associated with the x_{ki} and $\beta_{i'}$ are parameter vectors associated with the $x_{ki'}$.

The aforementioned maxdiff model was called a paired model in Flynn's study (Flynn et al., 2007). Moreover, another two sequential form models such as marginal model and sequential marginal model are also involved in profile case analysis. The marginal model assumes that decision makers choose the best and the worst options separately, unlike the paired model that considers the best-worst choice as a pair. Furthermore, the sequential marginal model is similar to the marginal model, since it assumes that decision makers might abandon the best (resp. the worst) option they initially chose from the attribute levels, and afterwards choose the worst (resp. the best) from the remaining attribute levels.

Case 3 (multi-profile case)

The multi-profile case or best-worst DCE (BWDCE) is the closest method to DCE, since it is designed to ask the respondent to choose the best and worst profiles in every choice set based on the DCE choice set (Lancsar et al., 2013). The potential best and worst options in a multi-profile case are all the profiles in the choice set X, when we assume there are n profiles ($n \geq 2$) in the choice set. Let $N = \{1, \ldots, n\}$ and profile k is the k th profile in the choice set X, $k = 1, \ldots, n$. The maxdiff model of best-worst choice probabilities in multi-profile case is presented as:

$$P_{BW}(ii' \mid X) = \exp(\mu(\sum_{k=1}^{n}(\beta_i x_{ki} - \beta_{i'} x_{ki'})) / \sum_{\substack{j,j'\in M \\ j'\neq j}} \exp(\mu(\sum_{k=1}^{n}(\beta_j x_{kj} - \beta_{j'} x_{kj'})) \qquad (7)$$

where x_{ki} is the attribute level in the profile that is chosen as the potential best option, and $x_{ki'}$, while $i \neq i'$ is the attribute level in the profile that is chosen as the worst option in profile k. μ represents a parameter that determines the scale of the utilities, while β_i are parameter vectors associated with the x_{ki} and $\beta_{i'}$ are parameter vectors associated with the $x_{ki'}$. Nevertheless, most applications of the multi-profile case use exploded logit to interpret the choice decision of respondents (Scarpa et al., 2011; Lancsar et al., 2013; Yoo et al., 2013; Gallego et al., 2015). In this study, we simply applied the sequential model as in the profile case for the sake of comparison.

3. Survey and data collection

3.1 Questionnaire

The questionnaire regarding the choice of mobile payments that was used in this study had three parts: In the first part, respondents are presented with the DCE choice set with an additional question (worst choice question) to obtain the DCE and BWS multi-profile data[2]. In each choice set, we presented three unlabeled profiles or alternatives: mobile payment A, mobile payment B, and mobile payment C. As presented in Table 1, each profile commonly includes six attributes (convenience, network externalities, transfer limitation, consumption limitation, probability of the password being deciphered, and cashback ratio) with each attribute having three levels. Most attributes regarding this question such as the items related to convenience, efficiency, limitless location, compatibility, perceived risk, perceived fee, network externalities, and promotional benefits were based on the studies of Dahlberg et al. (2008) and Madan and Yadav (2016). Additionally, we included cashback as an attribute. Furthermore, we used in the following analysis abbreviations in parentheses after descriptions of each level in Table 1 to be more concise. Moreover, we used Design-Expert Version 9 to create twenty-eight valid choice sets by employing the D-optimal design. Evidently, it was too cumbersome for respondents to answer all the choice sets. Thus, we further divided these choice sets randomly into four versions of questionnaires, and the respondents were only asked to answer one version that was randomly assigned to them.

[2] The BWS case 1 (i.e. object case) was not considered in this study, since it did not include the levels of attributes in the choice procedure.

Table 1. Attributes and their levels regarding mobile payments

Attributes	**Levels of attributes**
Convenience	Both internet access and entering the consumption amount of money are needed (*conv1*); Internet access is needed but entering the consumption amount of money is needless (*conv2*); Neither internet access nor entering the consumption amount of money is needed (*conv3*)

Network externalities	Accepted by 50% of merchants (*accep1*);
	Accepted by 75% of merchants (*accep2*);
	Accepted by 100% of merchants (*accep3*)

Transfer limitation	100,000 RMB / day (*trans1*); 200,000 RMB / day (*trans2*); 300,000 RMB / day (*trans3*)
Consumption limitation	1000 RMB / day (*cons1*); 3000 RMB / day (*cons2*); 5000 RMB / day (*con3*)
Probability of the password being deciphered	1% (*safety1*); 0.1% (*safety2*); 0.01% (*safety3*)
Cashback ratio	1% (*cashback1*); 5% (*cashback2*); 10% (*cashback3*)

The second section of the questionnaire targeted the BWS profile case to have a closer look at the preferences of respondents regarding the attribute levels of mobile payments, while the respondents were asked to choose the best and worst attribute levels from the profiles. Every choice set of the DCE or BWS multi-profile case had three profiles; therefore, there were twenty-one profiles in each version. Additionally, we randomly divided these profiles into three parts with seven profiles in each part and combined them with seven DCE or BWS multi-profile case choice sets[3]. The examples of DCE or BWS multi-profile case choice sets and BWS profile case choice sets are presented in Tables 2 and 3, respectively.

Questions in the third section of the questionnaire were related to demographic characteristics such as gender, academic degree, major, age, hometown, monthly living expenses, and monthly mobile payment consumption.

[3] In the survey, we did not separate the DCE choice sets from the BWS multi-profile case choice sets. The data for the DCE choice sets were selected from the profiles with the question “Please choose the mobile payment method you like the most.”

Table 2. An example of DCE or BWS multi-profile case choice sets

Attributes	**Mobile payment A**	**Mobile payment B**	**Mobile payment C**
Convenience	Both internet access and entering the consumption amount of money are needed	Internet access is needed but entering the consumption amount of money is needless	Neither internet access nor entering the consumption amount of money is needed
Network externalities	Accepted by 50% of merchants	Accepted by 75% of merchants	Accepted by 100% of merchants
Transfer limitation	100,000 RMB / day	200,000 RMB / day	300,000 RMB / day
Consumption limitation	1000 RMB / day	3000 RMB / day	5000 RMB / day
Probability of the password being deciphered	0.01%	0.1%	1%
Cashback ratio	10%	10%	5%
Please choose the mobile payment method you like the most			
Please choose the mobile payment method you like the least			

Table 3. An example of profile case choice sets

Best attribute level	**Mobile payment attribute levels**	**Worst attribute level**
	Both internet access and entering the consumption amount of money are needed	
	Accepted by 50% of merchants	
	100,000 RMB / day	
	1000 RMB / day	
	A 0.01% probability of the password being deciphered	
	10% Cashback	

Table 4. Demographic characteristics of the respondents (n=137)

Demographic characteristics	% in sample
Gender	
Male	42.34%
Female	57.66%
Age（mean=22）	
18-21	36.62%
22-25	61.19%
26 and above	2.19%
Hometown	
Urban area	32.85%
Rural area	67.15%
Academic degree	
Undergraduate student	40.88%
Postgraduate student	59.12%
Major in economics?	
Yes	54.74%
No	45.26%
Monthly living expenses (RMB)	
600-999	5.84%
1000-1499	29.93%
1500 and above	64.23%
Monthly ePay expenses (RMB)	
Below 599	3.65%
600-799	8.03%
800-999	22.63%
1000 and above	65.69%

$1≈6.4RMB

3.2 Data collection

Empirical data were collected using a local survey in Shanghai. The respondents were 137 undergraduate and graduate students of Shanghai University who had no technical background regarding mobile payment, though they had used it before. The demographic characteristics of participants are presented in Table 4. Female students accounted for 57.66% of the sample, and the average age was 22 years. Additionally, 67.15% came from rural areas, while 59.12% were postgraduate students. Moreover, 54.74% of the participants were students at the School of Economics, with most of the participants (64.23%) spending 1500 RMB or more on their monthly living expenses, excluding dormitory payments. Furthermore, most of the respondents (65.69%) spent more than 1000 RMB using mobile payment. Considering their living expenses, we can conclude that the amount spent for mobile payment accounts for most of their living expenses.

4. Estimation results and comparison between BWS and DCE

4.1 Estimation results in different models

Table 5 summarizes the conditional logit estimation results of the BWS profile case, DCE, and BWS multi-profile case. The models involved in the BWS profile case were paired model (PR), marginal model (MR), and sequential marginal model with the opposite order of the best and worst choices (SMR(best-worst) and SMR(worst-best)). Moreover, we simply evaluated the sequential order of the best and worst choices in the BWS multi-profile case (Case3(best-worst) and Case3(worst-best)) following Gallego et al. (2015).

The estimated results of these seven models were similar, with a few differences. The estimated parameters of *accep3*, *cons3*, *safety3*, and *cashback3* were positive and statistically significant in all the seven models. Thus, these results imply that the respondents considered the factors that mobile payment is accepted by 100% of merchants, the consumption limitation was 5000 RMB per day, the probability of the password being deciphered was 0.01%, and the cashback ratio was 10% as critical factors when compared with base levels. Conversely, the parameters of *trans3* in all the seven models were positive but not statistically significant, which indicate that the transfer limitation of 300,000 RMB/day had no no difference with 100,000 RMB/day limitation.

Regarding the differences among the estimation results in these seven models, the parameters of *conv2* were positive and statistically significant in the marginal model and both sequential marginal models in the BWS profile case. While the parameters of *conv3* were positive and statistically significant in the paired, marginal, and best-worst sequential marginal models in the BWS profile case. Additionally, the estimation results of *accep2* in the paired, marginal, and best-worst sequential marginal models in the profile case were negative and statistically significant. Nevertheless, the parameter of *accep2* in the DCE was positive and statistically significant, and the parameters of *trans2* in both the BWS multi-profile case models were negative and statistically significant. In contrast, the parameters of *cons2* were positive and statistically significant in all the BWS profile cases and DCE models, though not in two BWS multi-profile case models. In addition, no BWS profile case models showed statistically significant results for *safety2*; however, this variable showed significantly positive signs in the DCE and BWS multi-profile case models. Regarding *cashback2*, its parameters were not statistically significant in the paired model and two multi-profile case models, while they were positive and statistically significant in the rest four models at 1% level of confidence.

Furthermore, regarding the estimation results in all the seven models, we could not assert which levels of attributes had the most or least impact on respondents' choice. Although factors such as mobile payment being accepted by 100% of merchants, the consumption limitation of 5000 RMB per day, the probability of the password being deciphered being 0.01%, and the cashback ratio of 10% exhibited similar significant effects in all the models, there were obvious differences among estimated parameters. Without further investigation, we could not conclude which model could explain the data better and capture the respondents' decision more precisely. Therefore, we tried to find a statistical index to compare these models, which is presented in the next subsection.

Table 5. Conditional logit estimation results of the seven models

	PR	**MR**	**SMR(best-worst)**	**SMR(worst-best)**	**DCE**	**Case3(best-worst)**	**Case3(worst-best)**
Convenience (conv1 as base）							
conv2	0.127	0.265**	0.252**	0.267**	-0.011	0.017	0.015
	(1.24)	(2.47)	(2.33)	(2.51)	(-0.09)	(0.34)	(0.30)
conv3	0.398***	0.194*	0.188*	0.117	0.136	0.031	0.018
	(3.92)	(1.85)	(1.76)	(1.11)	(1.18)	(0.62)	(0.36)

Network externalities							
(accep1 as base)							
accep2	-0.218**	-0.241**	-0.263***	-0.156	1.302***	0.051	0.084
	(-2.32)	(-2.47)	(-2.69)	(-1.59)	(7.55)	(0.96)	(1.61)
accep3	2.246***	2.340***	2.301***	2.204***	2.014***	0.676***	0.692***
	(23.03)	(24.14)	(23.39)	(22.32)	(11.33)	(11.93)	(12.28)
Transfer limitation							
(trans1 as base)							
trans2	0.125	0.173	0.168	0.115	0.028	-0.131**	-0.131**
	(1.24)	(1.64)	(1.61)	(1.10)	(0.17)	(-2.34)	(-2.35)
trans3	0.074	0.016	0.010	0.079	0.164	0.090	0.083
	(0.72)	(0.14)	(0.09)	(0.73)	(1.01)	(1.52)	(1.41)
Consumption limitation							
(cons1 as base)							
cons2	0.497***	0.476***	0.457***	0.437***	0.321**	0.079	0.076
	(5.03)	(4.63)	(4.48)	(4.25)	(2.43)	(1.48)	(1.41)
cons3	0.497***	0.560***	0.534***	0.547***	0.306**	0.180***	0.163***
	(4.94)	(5.35)	(5.13)	(5.16)	(2.20)	(3.32)	(3.01)
Probability of the password being deciphered							
(safety1 as base)							
safety2	0.133	0.148	0.140	0.081	1.200***	0.131***	0.092*
	(1.39)	(1.51)	(1.41)	(0.81)	(9.13)	(2.68)	(1.85)
safety3	2.099***	2.264***	2.251***	2.254***	2.340***	1.029***	1.028***
	(21.14)	(23.22)	(22.71)	(22.47)	(16.70)	(17.85)	(17.56)
Cashback ratio							
(cashback1 as base)							
cashback2	0.136	0.282***	0.266***	0.260***	0.877***	0.054	0.017
	(1.43)	(2.88)	(2.65)	(2.69)	(5.80)	(0.90)	(0.29)
cashback3	1.657***	1.560***	1.532***	1.523***	1.998***	0.813***	0.804***
	(17.51)	(16.42)	(15.73)	(15.89)	(13.72)	(14.33)	(14.23)
Log-likelihood	-2362.964	-2456.677	-2383.239	-2403.475	-718.582	-1265.166	-1269.147
Observation	28,650	11,460	10,505	10,505	2,877	4,795	4,795
Count R square	0.279	0.523	0.526	0.526	0.674	0.701	0.691

Notes: ***, **, and * indicate statistically significant values at 1%, 5%, and 10% level of confidence, respectively. Z-statistics are reported in parentheses. PR: paired model; MR: marginal model; SMR(best-worst): best-worst sequential marginal model; SMR(worst-best): worst-best sequential marginal model; DCE: discrete choice experiment; Case3(best-worst): best-worst multi-profile case model; Case3(worst-best): worst-best multi-profile case model.

4.2 Discussion on the comparison among models

It was a strenuous task to find a statistical index to compare these models simultaneously, since the data coding method and the number of observations were different in the seven models. The traditional statistics used to calculate the goodness of fit usually relied on the log-likelihood values and number of observations such as McFadden's R square, AIC, BIC, etc. Nonetheless, the distinct log-likelihood values and number of observations in these models prevented us from comparing them by using the aforementioned statistics.

The count R square is not based on the log-likelihood values and number of observations, since it simply uses the ratio of the correct prediction obtained from the regression model to the total number of observations. To calculate this index, we should first define what the correct prediction is. The count R square is usually utilized to calculate the goodness of fit in the logit model. Assuming the observed choice y to be 0 or 1 and the predicted probability $\pi_i = \hat{\Pr}(y=1 \mid x_i)$, we define the predicted outcome as follows:

$$\hat{y}_i = \begin{cases} 0 & if\ \hat{\pi}_i \le 0.5 \\ 1 & if\ \hat{\pi}_i > 0.5 \end{cases} \tag{7}$$

where if the predicted probability is less than or equal to 0.5, the predicted outcome is noted as 0; while if the predicted probability is higher than 0.5, the predicted outcome is noted as 1. The correct prediction is defined as the number of observed choices equals the number of predicted outcomes. Therefore, the formula of count R square could be presented as follow:

$$R^2_{Count} = \frac{1}{N} \sum_j n_{jj} \tag{8}$$

where n_{jj} s are the number of correct predictions for outcome j, and N is the total number of observations (Long and Freese, 2000). However, this definition might be inappropriate in the conditional logit model, since all the predicted probabilities in one choice set might be less than 0.5. Hence, Long and Freese (2000) calculated count R square in conditional logit model using another criterion. They redefined the predicted outcome as follows:

$$\hat{y}_{ik} = \begin{cases} 0 & if\ \hat{\pi}_{ik} \neq \max\{\pi_{ik}\} \\ 1 & if\ \hat{\pi}_{ik} = \max\{\pi_{ik}\} \end{cases} \tag{9}$$

where $\hat{y}_{ik}$ is the predicted outcome for i th level or profile in choice set k , $\hat{\pi}_{ik}$ is the predicted probability for i th level or profile in choice set k. The predicted outcome is equal to 1, if the predicted probability is the maximum probability in choice set k , else the predicted outcome would be equal to 0.[4] Based on this definition, they calculated the count R square of conditional logit model with the only focus on the part that predicted outcomes are 1, which was similar to the concept called sensitivity in classification table (Hosmer Jr. et al., 2013)[5]. The formula of count R square in conditional logit could be summarized as follows:

$$\tilde{R}^2_{Count} = \frac{n_{11}}{n_{11} + n_{01}} \tag{10}$$

where n_{11} is the number of correct predictions for both observed choice and predicted outcome equaling 1, while n_{01} is the number of the wrong prediction for observed choice being 0, but the predicted outcome being 1.

The count R squares of the seven models are presented in the last row of Table 5. The results showed that the highest count R square appeared in the BWS best-worst multi-profile case model, while the paired model in BWS profile case had the lowest count R square. As a whole, the count R squares in BWS profile case models were less than that in DCE, whereas those in BWS multi-profile case models were the highest among these three methods. Additionally, in the profile case, the count R square in the paired model was lower than other marginal models, while the marginal model's count R square was less than that in both sequential marginal models. Moreover, in the multi-profile case, the best-worst multi-profile case model had

[4] In the BWS paired model, if the respondent chose the specific combination of best and worst items, the choice variables were coded as 1, else 0; and in the DCE, the choices of the best item were coded as 1, else 0. However, in other BWS case models, the best and worst options were treated as being separated, the chosen alternatives in these models needed to be coded as 1 twice (i.e., the best choice and the worst choice). Hence, the predicted outcome in paired model and DCE could follow the definition in Equation (9), while other five models needed to code the best (resp. worst) choice after their worst (resp. best) choice.

[5] There are two reasons why we chose the predicted outcomes of which the values were 1. Firstly, the empirical methods were inclined to discover what respondents chose rather than what they did not choose. Secondly, all seven models had distinct choice options that induced more zeros predicted outcomes to be generated in some models such as the paired model. Therefore, the extra zeros would bring extra noise, which would cause a comparison error.

higher count R square than the worst-best one, though the difference was not large. From the aforementioned comparison, we could conclude that the BWS best-worst multi-profile has the best goodness of fit among these seven models, which seems appropriate, since this model contains the most information.

Evidently, all the BWS profile case models contain less information than the DCE and the BWS multi-profile case models; therefore, they show a lower goodness of fit than the rest two methods. Moreover, mental process in the paired model is complicated, which might generate inaccurate predictions in addition to worsening the goodness of fit. According to our analysis, the BWS best-worst multi-profile case model could provide a better prediction for respondents' behavior.

5. A comparison between LCM and MLM

5.1 Estimation of MLM and LCM in the BWS best-worst multi-profile case

Table 6 summarizes the estimation results with the specifications of Latent Class Model (LCM) and Mixed Logit Model (MLM) in the BWS best-worst multi-profile case. Additionally, we define all the variables in MLM as random variables and the standard deviations are presented in the third column of Table 6. Demographic characteristics of the respondents were utilized to select an optimal number of latent classes in LCM, with two latent classes emerging as dominant.

Table 6. Estimation results of LCM and MLM in the BWS best-worst multi-profile case

	MLM	Standard Deviation	LCM	
			Class 1	Class 2
Convenience *（conv1 as based）*				
conv2	-0.054	-0.252**	0.111	-0.164**
	(-0.70)	(-2.07)	(1.44)	(-1.96)
conv3	0.139	0.824***	0.132	-0.066
	(1.29)	(7.24)	(1.55)	(-0.67)
Network externalities *(accep1 as based)*				

accep2	0.141*	-0.239*	0.164**	-0.013
	(1.79)	(-1.77)	(2.03)	(-0.14)
accep3	1.242***	1.053***	0.945***	0.557***
	(9.08)	(8.32)	(9.86)	(5.73)
Transfer limitation (trans1 as based)				
trans2	-0.097	0.286*	-0.212**	0.099
	(-1.12)	(1.71)	(-2.28)	(1.02)
trans3	0.135	0.087	0.152*	-0.009
	(1.54)	(0.45)	(1.66)	(-0.09)
Consumption limitation (cons1 as based)				
cons2	0.178**	-0.045	0.186**	0.002
	(2.22)	(-0.24)	(2.24)	(0.02)
cons3	0.253**	0.676***	0.048	0.368***
	(2.55)	(5.30)	(0.59)	(3.46)
Probability of the password being deciphered (safety1 as based)				
safety2	0.264***	0.431***	0.131*	0.151*
	(3.16)	(4.24)	(1.72)	(1.75)
safety3	1.818***	1.455***	0.403***	1.901***
	(10.94)	(7.85)	(4.44)	(14.80)
Cashback ratio (cashback1 as based)				
cashback2	0.010	0.315**	-0.022	0.034
	(0.11)	(2.44)	(-0.24)	(0.34)
cashback3	1.374***	0.710***	0.842***	1.008***
	(11.52)	(3.85)	(8.86)	(9.84)
Log-likelihood	-1121.872		-1180.788	
Observations	4,795		4,795	

Notes: ***, **, and * indicate statistically significant values at 1%, 5%, and 10% level of confidence, respectively. Z-statistics are reported in parentheses.

The results in the MLM and LCM model were slightly different from those in the conditional logit model (see the second column from the right side in Table 5). We found that the parameters of *conv3* and *cashback2* in all three models (i.e., conditional logit model, MLM and LCM) were not statistically significant, while the parameters of *accep3, safety2, safety3,* and *cashback3* were statistically significant. However, there were some differences among the regression results of these three models.

Additionally, the parameter of *conv2* was negative and statistically significant only in Class 2 of LCM, though not in the conditional logit model and MLM. Compared with these in the conditional logit model, the parameters of *accep2* and *cons2* were statistically significant in both MLM and LCM. Furthermore, the parameters of *trans2* were negative and statistically significant in conditional logit model and Class 1 of LCM, though not in MLM. Finally, the parameters of *cons3* were statistically significant in all models except for Class 1 of LCM.

5.2. A comparison between MLM and LCM

In this subsection, we applied three tests to non-nested models (i.e., the Akaike index, Vuong test, and distribution free test) to make a comparison between MLM and LCM in the BWS best- worst multi-profile case.

5.2.1. The Akaike index

Ben-Akiva and Swait (1986) proposed an index based on the Akaike Information Criterion (AIC) to implement a test for selecting the superior model. Suppose there are two non-nested models 1 and 2, Model 1 includes K_1 independent variables, while model 2 includes K_2 independent variables. Assume $K_1 \geq K_2$ and either the two models have different functional forms or the two sets of variables are different by at least one element. Therefore, the Akaike likelihood ratio index called by Ben-Akiva and Swait (1986) could be used to measure goodness of fit as follows:

$$\rho_j^2 = 1 - \frac{L_j - K_j}{L(0)} \tag{11}$$

where L_j is the log likelihood at convergence for model j, K_j is the number of independent variables in model j, and the $L(0)$ is the log likelihood for constant only (Shen, 2006). The null hypothesis is proposed assuming model 2 as the true model; thus, the probability of the index for model 1 will be greater than that of model 2 is asymptotically bounded by a function given in Equation (12):

$$\Pr(\left|\rho_2^2 - \rho_1^2\right| \geq Z) \leq \Phi(-\sqrt{-2ZL(0) + (K_1 - K_2)} \qquad (12)$$

where Z is the difference of the index between model 1 and model 2 and is assumed to be larger than zero, while Φ is the standard normal cumulative distribution function. Additionally, equation (12) sets an upper bound for the probability that model 1 is incorrectly selected as the true model when model 2 is the true model.

Regarding the aforementioned definition, we calculated the probability in Equation (12) for the best-worst multi-profile case and assumed LCM was model 1 and MLM was model 2. The upper bound of the probability was $P \leq \Phi(-10.855) \approx 0$, which means that MLM is superior to the LCM.

5.2.2. Vuong test

Vuong test is another non-nested test that was proposed by Vuong (1989). This test is based on a likelihood ratio statistic called Kullback-Leibler Information Criterion (KLIC) (Kullback and Leibler, 1951). The KLIC measures the distance between the true model and a hypothesized model regarding the likelihood function. Formally, the KLIC could be written as follows:

$$KLIC = E[\ln h(Y \mid X, \alpha)] - E[\ln f(Y \mid X, \beta)] \qquad (13)$$

where $h(Y \mid X, \alpha)$ is the true model, while $f(Y \mid X, \beta)$ is the hypothesized model, and E is the expectation under the true distribution. The logic of Vuong test is that if one of the competing models is closer to the true model based on KLIC, then it will surpass others.

Additionally, the Vuong test considers the average difference in the log-likelihoods of two competing statistical models, with the null hypothesis of the test is that this average difference is zero (Clarke and Signorino, 2010). We denoted f as model 1 with covariates X and coefficient β, and denoted g as model 2 that includes the covariates Z and the coefficient γ. Hence, the null hypothesis could be written as:

$$H_0: \; E\left[\ln\frac{f(Y\mid X,\beta)}{g(Y\mid Z,\gamma)}\right]=0 \tag{14}$$

The null hypothesis suggests that the two models are equally close to the true model under a general condition that:

$$\frac{1}{n}LR(\hat{\beta},\hat{\gamma})\xrightarrow{a.s.}E\left[\ln\frac{f(Y\mid X,\beta)}{g(Y\mid Z,\gamma)}\right] \tag{15}$$

where $LR(\hat{\beta},\hat{\gamma})$ is the estimated difference in the log-likelihoods of the two models (i.e., $L_f(\hat{\beta})-L_g(\hat{\gamma})$). Afterwards, we normalized the log-likelihood ratio statistic. Subsequently, we found that the Vuong test statistic V was normally distributed under null hypothesis as follows：

$$V=\frac{LR(\hat{\beta},\hat{\gamma})}{(\sqrt{n})\hat{s}}\xrightarrow{D}N(0,1) \tag{16}$$

where $\hat{s}=\frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{f(Y_i\mid X_i,\hat{\beta})}{g(Y_i\mid Z_i,\hat{\gamma})}-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(Y_i\mid X_i,\hat{\beta})}{g(Y_i\mid Z_i,\hat{\gamma})}\right]^2$

The selection of models followed the criteria below:

(1) Under the hypothesis that the models are “equivalent”, $V\xrightarrow{D}N(0,1)$.

(2) Under the hypothesis that $f(Y\mid X,\beta)$ is “better”, $V\xrightarrow{a.s.}+\infty$.

(3) Under the hypothesis that $g(Y\mid Z,\gamma)$ is “better”, $V\xrightarrow{a.s.}-\infty$.

In fact, if V is a value between -1.96 and 1.96, we cannot decide which model is superior. Whereas if V is larger than 1.96, $f(Y\mid X,\beta)$ is superior; and if V is less than -1.96, $g(Y\mid Z,\gamma)$ is superior. We assumed $f(Y\mid X,\beta)$ was LCM, and $g(Y\mid Z,\gamma)$ was MLM in our BWS best-worst multi-profile case. The Vuong test statistic $V=-19.900<-1.96$, which means that MLM is superior to LCM.

5.2.3. Distribution free test

The distribution free test is also based on KLIC; however, it considers the median difference in the log-likelihoods of two competing statistical models (Clarke and Signorino, 2010). The null hypothesis is that if two models are equally close to the truth, half of the log-likelihood ratios should be greater than zero; therefore, the null hypothesis could be formally presented as follows:

$$H_0 : \Pr\left[\ln \frac{f(Y \mid X, \beta)}{g(\mathrm{Y} \mid Z, \gamma)} > 0 \right] = 0.5 \qquad (17)$$

If we denote $d_i = \ln f(Y \mid X, \hat{\beta}) - \ln g(Y \mid Z, \hat{\gamma})$, then the test statistic could be written as:

$$B = \sum_{i=1}^{n} I_{(0,+\infty)}(d_i) \qquad (18)$$

where I is the indicator function that calculates the number of positive sign in differences (i.e., d_i) and it is distributed Binomial with parameters n and $p = 0.5$. Thus, if the models are equally close to the truth, half of the individual log-likelihood ratios should be greater than zero and the other half should be less than zero. If model f is "better" than model g, then more than half of the individual log-likelihood ratios should be greater than zero. Conversely, if model g is"better" than model f, then more than half of the individual log-likelihood ratios should be less than zero (Clarke and Signorino, 2010). The f is assumed as LCM, and g is assumed as MLM in our BWS best-worst multi-profile case. Subsequently, we calculated the distribution free test statistic $B = 2300 < \frac{n}{2} \approx 2398$, which means MLM is superior to LCM.

According to the aforementioned three non-nested model tests, we conclude that the MLM is dominant in the BWS best-worst multi-profile case. Additionally, we ran the three non-nested tests on MLM and LCM in the other six models. The results presented in Table 7 suggest that MLM is superior to LCM in all the models. While our results of comparing MLM with LCM are inconsistent with these in the previous DCE studies (Greene and Hensher, 2003; Shen, 2009, the econometric model selection in BWS should be prudent, since the results of the model selection might vary when using different data in various circumstances.

Table 7. Results of the three non-nested tests on MLM and LCM in the seven models

	Akaike index (Φ)	**Vuong test (V)**	**Distribution free test (B)**
PR	-11.337	-71.620	9693(14325)
MR	-14.156	-30.249	4490(5730)
SMR(best-worst)	-13.160	-25.313	4180(5252.5)
SMR(worst-best)	-12.351	-21.615	4204(5252.5)
DCE	-6.238	-16.687	1342(1438.5)
Case 3(best-worst)	-10.855	-19.900	2300(2397.5)
Case 3(worst-best)	-10.448	-18.347	2291(2397.5)

Note: The values of $\frac{n}{2}$ in the seven models are listed in parentheses after the distribution free test statistic B .

6. Conclusion

In this study, we applied a survey data to consumers' preference regarding mobile payment in Shanghai to compare different methods of the BWS profile case, DCE, and BWS multi-profile case. Seven models were regressed and the estimated parameters could not show clearly which method is better. Therefore, the statistical index of count R square was utilized to compare the goodness of fit among these seven models. The results exhibited that the BWS best-worst multi-profile case model had the best goodness of fit, whereas the BWS paired model in profile case had the worst goodness of fit. Additionally, we implemented three non-nested model tests to compare MLM and LCM specifications. The results were robust in all three tests that MLM was superior to LCM in all the cases.

Finally, there are two important implications related to future research. Firstly, the comparisons among the BWS profile case, DCE, and BWS multi-profile case are not enough to provide clear concepts in the existing literature. Therefore, this issue should be further investigated. Additionally, more statistical indices that can help researchers compare these methods should be explored in future studies. Secondly, although our study suggests that the MLM is superior to LCM in the BWS method, the results might vary when using different data. Hence, further studies regarding the comparison between these two specifications in both the DCE and BWS methods will be greatly beneficial.

Acknowledgements

Financial support from the Japanese Ministry of Education, Culture, Sports, Science and Technology through Grants-in-aid for Scientific Research (C) 17885015, 19115990, and 19187135 are gratefully acknowledged. All the views expressed in this paper and any errors are the sole responsibility of the authors.

References

Ben-Akiva, M., Lerman, S. R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press.

Ben-Akiva, M., Swait, J., 1986. The Akaike likelihood ratio index. *Transportation Science*, *20*(2), 133-136.

Clarke, K. A., Signorino, C. S., 2010. Discriminating methods: Tests for non-nested discrete choice models. *Political Studies*, *58*(2), 368-388.

Dahlberg, T., Mallat, N., Ondrus, J., Zmijewska, A., 2008. Past, present and future of mobile payments research: A literature review. *Electronic Commerce Research and Applications*, *7*(2), 165-181.

Finn, A., Louviere, J. J., 1992. Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing*, *11*(2), 12-25.

Flynn, T. N., Louviere, J. J., Peters, T. J., Coast, J., 2007. Best–worst scaling: What it can do for health care research and how to do it? *Journal of Health Economics*, *26*(1), 171-189.

Flynn, T. N., 2010. Valuing citizen and patient preferences in health: Recent developments in three types of best–worst scaling. *Expert Review of Pharmacoeconomics & Outcomes Research*, *10*(3), 259-267.

Flynn, T. N., Peters, T. J., Coast, J., 2013. Quantifying response shift or adaptation effects in quality of life by synthesising best-worst scaling and discrete choice data. *Journal of Choice Modelling*, *6*, 34-43.

Greene, W. H., Hensher, D. A., 2003. A latent class model for discrete choice analysis: Contrasts with mixed logit. *Transportation Research Part B: Methodological*, *37*(8), 681-698.

Gallego, G., Dew, A., Lincoln, M., Bundy, A., Chedid, R. J., Bulkeley, K., Brentnall， J., Veitch, C., 2015. Should I stay or should I go? Exploring the job preferences of allied health professionals working with people with disability in rural Australia. *Human Resources for Health*, *13*(1), 53.

Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X., 2013. *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86.

Lancsar, E., Louviere, J., Donaldson, C., Currie, G., Burgess, L., 2013. Best worst discrete choice experiments in health: Methods and an application. *Social Science & Medicine*, *76*, 74-82.

Long, J. S., Freese, J., 2000. Scalar measures of fit for regression models. *Stata Technical Bulletin*, *56*, 34-40.

Louviere, J. J., Woodworth, G., 1983. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, *20*(4), 350-367.

Louviere, J. J., Flynn, T. N., Marley, A. A. J., 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Madan, K., Yadav, R., 2016. Behavioural intention to adopt mobile wallet: a developing country perspective. *Journal of Indian Business Research*, 8(3), 227-244.

Potoglou, D., Burge, P., Flynn, T., Netten, A., Malley, J., Forder, J., Brazier, J. E., 2011. Best–worst scaling vs. discrete choice experiments: An empirical comparison using social care data. *Social Science & Medicine*, *72*(10), 1717-1727.

Scarpa, R., Notaro, S., Louviere, J., Raffaelli, R., 2011. Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in alpine grazing commons. *American Journal of Agricultural Economics*, *93*(3), 813-828.

Shen, J., 2006. A review of stated choice method. *International Public Policy Studies*, *10*(2), 97-121.

Thurstone, L. L., 1927. A law of comparative judgment. *Psychological Review*, *34*(4), 273.

Vuong, Q. H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307-333.

Whitty, J. A., Ratcliffe, J., Chen, G., Scuffham, P. A., 2014. 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.

Xie, F., Pullenayegum, E., Gaebel, K., Oppe, M., Krabbe, P. F., 2014. Eliciting preferences to the EQ-5D-5L health states: Discrete choice experiment or multiprofile case of best–worst scaling? *The European Journal of Health Economics*, *15*(3), 281-288.

Yoo, H. I., Doiron, D., 2013. The use of alternative preference elicitation methods in complex discrete choice experiments. *Journal of Health Economics*, *32*(6), 1166-1179.