Published in "Journal of business research", 2022, vol. 148, pp. 368-377, which should be cted to refer to this work.
DOI: 10.1016/j.jbusres.2022.04.064

A Comparison of Best-Worst Scaling and Likert Scale Methods on Peer-to-Peer Accommodation Attributes

A Research Paper Published at
Journal of Business Research

Special Issue on Scale Development in Tourism and Leisure Research: New Approaches, Theoretical and Methodological Issues.

Cindy Yoonjoung Heo
Associate Professor
EHL Hospitality Business School
HES-SO// The University of Applied Sciences and Arts of Western Switzerland, Switzerland
Phone: (41) 21 785 1589
E-mail: cindy.heo@ehl.ch

Bona Kim
Assistant Professor
School of Global Convergence Studies
Inha University, Korea
Phone: (82) 32-860-9122
E-mail: bona.kim@inha.ac.kr

Kwangsoo Park
Associate Professor
Department of Apparel, Design, and Hospitality Management
North Dakota State University, USA
Phone: (1) 701-231-7355
E-Mail: kwangsoo.park@ndsu.edu

Robin M. Back
Associate Professor
Rosen College of Hospitality Management
University of Central Florida, USA
Phone: (1) 407-903-8207
E-Mail: robin.back@ucf.edu

A Comparison of Best-Worst Scaling and Likert Scale Methods on Peer-to-Peer Accommodation Attributes

ABSTRACT

Surveys based on Likert scales continue to dominate market research practice despite their limitations. Several researchers have suggested adopting different types of scales and a unique alternative for rating the importance level of several attributes is Best−Worst Scaling (BWS). The purpose of this study is to compare two scaling approaches, the Best-Worst Scale (BWS) and the Likert Scale to explore their advantages and disadvantages. This study tried to identify the relative importance of Peer-to-Peer (P2P) accommodation attributes using the aforementioned two scaling approaches. A comparison of the results found that the BWS approach helps to validate priorities from a customer perspective by achieving better discrimination among attributes, while the Likert scale approach is useful for comparing group differences such as gender differences.

Keywords: Best−Worst scaling, Likert scale, scale comparison, P2P accommodation attributes

1. Introduction

The issue of scale is important to building knowledge in social science research because it is the process of measuring qualitative or quantitative attributes of entities. Gibson et al. (2000) defined scale as “the spatial, temporal, quantitative, or analytical dimensions used to measure and study any phenomenon” (p. 218). Although relatively little research has been dedicated to scale issues in the social sciences when compared to the natural sciences (Gibson et al, 2000), social science research, including the tourism and hospitality field, is continuously calling for scale development. This attempt supports rigorous research practices in measuring phenomena of interest represented as constructs such as individuals' perceptions, opinions, or preferences (Joshi et al., 2015).

The constructs are expressed by multiple manifested items in questionnaires and measured by psychometric tools such as the Likert scale and rating scales. These conventional scales are most frequently adopted by researchers (Bertram, 2007). However, it has been a challenge to enhance methodological advances that can increase reliability of measurement and statistical powers (Burton et al., 2021) and there is, thus, a necessity for improving the robustness of research measurement scales (Burton et al., 2021; Kiritchenko & Mohammad, 2017).

One of the most adopted scales in social science studies is the Likert scale, where responses to questions are measured on a continuum of two endpoints (Dittrich et al., 2007). The Likert scale is an interval scale assuming that two consecutive points are reflected within equal distance in variation (Crask & Fox, 1987). The Likert scale has long been considered a convenient scale for obtaining participants' preferences or degree of agreement with a set of statements, constructing and modifying responses, and generating appropriate results for statistical inference (Bertram, 2007; Li, 2013).

Despite the Likert scale being considered a convenient scale, researchers have pointed out inherent limitations associated with it and have claimed that the Likert scale is not sufficiently reliable (e.g., Chrzan & Skrapits, 1996; Cohen & Markowitz, 2002; Cohen & Neira, 2003; Louviere et al., 1995). One of the issues is that the Likert scale is a non-comparative scaling technique that measures a single trait at a time as a unidimensional tool so that it does not reflect the complexity of human opinions (Bertram, 2007; Joshi et al., 2015). Accordingly, it has been argued that the Likert scale may not be the best scale to measure the importance level among various attributes. Several researchers have suggested adopting different types of scales (Cohen, 2009; Li, 2013) and a unique alternative for rating the importance level of several attributes is Best−Worst Scaling (BWS), introduced by Finn and Louviere (1992).

BWS is a theory-based scaling method that applies a discrete choice experiment based on a random utility theory (Flynn & Marley, 2014). The discrete choice-based evaluation considers how people evaluate attributes as top and bottom in a list (Flynn & Marley, 2014). A type of BWS, the BWS object case, includes a series of choice tasks, each of which contains a different set of items. In each choice situation, respondents are asked to choose the "best" and the "worst" item (e.g., "most important" and "least important" or "most useful" and "least useful") from a subgroup of items derived from a list (Louviere & Islam, 2008).

Some scholars argued that BWS helps to avoid reliability issues of conventional scales and is a suitable approach to identify the relative values of complex subjects. When the list of items of interest to the researcher is long and respondents indicate that all items are quite important, the results may not be very meaningful. Indeed, several studies have shown that the BWS is superior to rating scales (Lee et al., 2007) and is not vulnerable to problems such as different response styles of respondents (Baumgartner & Steenkamp, 2001). BWS has been fairly popular in the past but interest has declined in recent years. Perhaps part of the reason for this decline is that researchers lack the knowledge on how to proceed with BWS or do not fully appreciate the merits of BWS. Therefore, the objective of this study is to identify the strengths and weaknesses of the two scaling approaches and to provide guidelines for researchers for future applications.

To compare the two scaling approaches (i.e., BWS and Likert scales), the relative importance of various peer-to-peer (P2P) accommodation attributes is measured using each of the approaches. Today's tourists are easily overwhelmed by an enormous number of options from which to choose as well as their complexity, due to the numerous information channels and online platforms that have become available. The emergence of peer-to-peer (P2P) accommodation platforms, such as Airbnb, created an alternative lodging option and added yet another layer of complexity. In general, consumers' accommodation choice is influenced by a variety of factors, such as the various accommodation offerings and personal preferences (Chu & Choi, 2000; Kim et al., 2019). While traditional hotel attributes and services are rather standardized, all P2P accommodation options are unique and different. In addition, previous literature has shown that the attributes of P2P accommodation hosts are just as important as the attributes of the property (e.g., Chattopadhyay & Mitra, 2020; Ma et al., 2017).

Accordingly, the range of attributes of P2P property and host is very broad and intricate, and it is not easy for service providers (i.e., accommodation hosts) to understand the salient attributes that are valued by their customers. We believe that the variety and complexity of amenities of P2P accommodations can highlight the different aspects of the two scaling methods. Therefore, this study tried to find key attributes of P2P accommodation in terms of property and host by comparing the results using the previously mentioned two scaling approaches. Methodological contribution and practical advice for future research are discussed based on the findings of this study.

2. Literature Review

2.1 Likert Scales

The Likert scale was introduced by Rensis Likert in 1932 and has been widely used to measure observable attributes in social science studies (Li, 2013). It is used to indicate subjects' level of agreement on x-point Likert scales or to rate the importance level of topic attributes beliefs and opinions. (Chu & Choi, 2000; Qu et al., 2000). Questionnaires based on the Likert scale allow respondents to respond in a degree of agreement instead of forcing them to take a stand on a particular topic. The typical scale used in marketing normally labels each scale category with adjectival descriptors, such as "important" or "not important", "good" or "bad" (Cohen, 2009). Respondents can easily understand and answer the questions based on the Likert scales and responses are easy to code when accumulating data. Furthermore, it is convenient for constructing and modifying responses, generating appropriate results for statistical inference with good reliability, and facilitating different data analysis methods for a large quantity of data with little time and effort (Li, 2013). Therefore, it is most commonly used for scaling responses in survey research as an efficient and inexpensive method of data collection.

However, a few scholars have discussed the limitation of the Likert scales. Joshi et al. (2015) discussed controversies regarding the analysis and inclusion of points on Likert scales. Although the Likert scale was proposed as an interval scale by assuming that two consecutive points are reflected within equal distance in variation, respondents may not equivalently recognize the distances between two points of the scale (Crask & Fox, 1987). Interpretation can be problematic when ordinal data are used in statistical analyses that require interval scale variables (Harwell & Gatti, 2001). Further, the results of ratings from the scale may have different implications for different individuals. Another issue is related to individuals' tendency to avoid selecting the extreme options on the scale. Even though an extreme choice would be the most accurate, respondents may avoid choosing the extreme options because of the negative implications involved with extremists. Furthermore, several studies found that cultural and ethnic groups differ in their extreme response style (e.g., Hui & Triandis, 1989). On the other hand, Garrido et al. (2013) pointed out Cronbach's alpha, which most of the studies based on Likert data have been using, and which has often misinterpreted and/or misused. Criticism about the Likert scale leads several researchers to suggest different types of scales.

2.2. Best−Worst Scaling (BWS)

The BWS (also known as maximum difference scaling) was proposed by Louviere and Woodworth (1983) as one of the scales for rating the importance level of several attributes. The BWS is based on Thurstone's (1927) random utility theory for paired comparisons (Cohen, 2009). The BWS requires subjects to select only one most and one least preferred item in each choice set while considering trade-offs between items (Cohen, 2003). Given this, BWS identifies the rank among items that have subtle weights on importance without any bias resulting from cultural differences (Lusk & Briggeman, 2009). The statistical model underlying BWS is that the relative choice probability of a specified pair relates to the distance between the two attribute levels on the latent utility scale (Flynn et al., 2007). It is assumed that each individual recognizes good or bad, and even best or worst as the extreme levels in placing importance on an item (Finn & Louviere, 1992). Individuals are asked to choose the best item and the worst item in each choice set and the farthest distance between the best and worst items on an underlying latent scale indicates the degree of importance (Cohen, 2009). By considering the distances for two items, the relative importance of items can be determined in consideration of the benefits of trade-offs (Louviere & Islam, 2008).

BWS has been adopted to identify preferred or important items related to attributes, benefits, or characteristics of consumers in wine and food-related studies (e.g., Cohen, 2009; Goodman et al., 2005; Lockshin et al., 2011; Lockshin et al., 2017; Lusk & Briggeman, 2009) and in hotel and tourism studies (e.g., Kim et al., 2019; Scarpa et al., 2011). A few studies have compared importance weights by adopting multiple scaling methods categorized as indirect or direct scales to address the validity issue of the BWS method, (e.g., Jaeger & Cardello, 2009; Lagerkvist, 2013; Louviere & Islam, 2008; Mueller et al., 2009). The summary of previous studies on BWS method is presented in Appendix 1.

In summary, previous studies comparing BWS and other scales have focused mostly on the efficiency and effectiveness of scales, from the perspectives of practitioners or researchers (Jaeger & Cardello, 2009). Also, they mainly highlighted technical aspects of scales, such as ease and commonality of use, sensitivity, or shortcomings. However, few have compared the actual results generated by using them. Since BWS is a useful tool for identifying importance levels among factors by comparing the subtle discriminations on importance weights (Kim et al., 2019), the attributes related to P2P accommodation sharing as a current emergent issue in the tourism industry is the focus of the present study. In addition, gender differences in the perceived importance of Airbnb accommodation attributes were explored to highlight the different aspects of two scaling methods.

3. Methodology

3.1. Study Design

This study was designed to compare two scaling approaches – the Likert scale and the BWS – in identifying the importance levels of host and accommodation-related attributes of P2P accommodation. To adopt both scales, a procedure involving multiple steps was followed. In the initial step, unique attributes of P2P accommodation sharing were explored in a twofold manner. The literature on P2P accommodation attributes was extensively reviewed to list a variety of relevant attributes. Through the review process, we discovered that 'accommodation host' is a distinct element in the P2P transaction context unlike traditional hotels (e.g., Edelman & Luca, 2014; Ert et al., 2016; Liang et al., 2018; Ma et al., 2017; Wang & Nicolau, 2017). Therefore, this study separated those attributes associated with hosts from accommodation-related attributes. Next, based on the two lists of attributes, we conducted six interviews with Airbnb users to finalize the salient P2P accommodation attributes. The survey questionnaire for the Likert scales was developed based on 13 host-related and 13 accommodation-related attributes for P2P accommodation sharing and the list is presented in Table 1.

[Table 1]

In the second step, the design of 'choice sets' for the BWS approach should include all items over an equal number of times for all possible comparisons based on the multinomial logit model (Louviere & Woodworth, 1983). In this study, Balanced Incomplete Block Design (BIBD) was adopted since it is the most common design to conduct counting-based analyses for organizing a series of choice sets (Auger et al., 2007; Cohen, 2009; Goodman et al., 2005; Louviere & Woodworth, 1983). This ensures the constant occurrence and co-occurrences of items in a set of choices and minimizes the chance that respondents may make unintended assumptions about the items as a type of choice set design (Flynn & Marley, 2014). A useful feature of BIBD is to ensure that every item appears in every possible position the same number of times while minimizing the number of subsets including a certain number of items (Louviere & Woodworth, 1983). It is regarded as an extension of paired comparison and represents the most robust defense against any inclination of respondents to read too much into the size or composition of the choice sets (Flynn & Marley, 2014). BIBD is based on a Latin Square design with *n* items arranged by *n* rows and *n* columns. The items for each row and column are in different positions and are indicative of a block or a choice set (Weller & Romney 1988). This method allows many items to be compared to obtain the full rank of all items in a small number of subsets (Auger et al., 2007; Cohen, 2009).

By adopting the BIBD, we designed choice sets consisting of items that occur in every possible subset the same number of times in the same number of choice sets. The possible combinations of BIBD for 13 attributes were arranged as (v, k, r, b) where v is the treatment, k the number of items in each choice set, r the repetition per level, and b the number of choice sets insofar as k < v. According to the study by Yasmin et al. (2015), the major conditions for a BIBD are:

1) $r = b\ k/v$,

2) treatment does not appear more than once in any choice set, and

3) as λ is the pair frequency, all unordered pairs of attributes appear exactly in λ blocks (1), where $\lambda = r\ (k-1) / (v-1) = b\ k\ (k-1) / v\ (v-1)$ is often referred as the concurrence parameter of a BIBD.

Given the condition, combinations of (13, 3, 6, and 26) are generated for the BIBD of this study. In each choice set including three attributes, the Best/most important attribute and the Worst/least important attributes are selected in 26 choice sets.

The questionnaire was composed of four sections including a screening question, BWS questions, Likert scale questions, and demographic profile questions. A screening question, whether participants have stayed at a P2P accommodation during the past five years, was asked to meet the appropriate sample criteria. The questionnaire began with each of the 26 choice sets including three attributes for both host and accommodation sections, respectively. A total of 52 choice sets (26 choice sets for host-related attributes and 26 choice sets for accommodation-related attributes of P2P accommodation) were asked, where respondents were asked to select each attribute as the MOST important, and an attribute as the LEAST important among three attributes in a choice set. For the questions using the Likert scale, the host- and accommodation-related attributes were designed using a 5-point Likert scale.

3.2. Data Collection

The population of interest was P2P accommodation users in the United States. The survey was administered in August 2017 and collected 304 responses; however, a total of 302 responses were used for data analysis, with two inappropriate responses deleted. To conduct the survey, an online survey platform − Qualtrics − was used. Voluntary respondents were recruited online using Amazon's Mechanical Turk (MTurk) crowdsourcing platform. MTurk is a platform used by researchers to recruit subjects to complete Human Intelligence Tasks (HITs) (Strich et al., 2017). MTurk has been used increasingly for surveys in the social sciences (e.g., Aquinis et al., 2021; Arceneaux, 2012; Berinsky et al., 2014; Healy & Lenz, 2014) and is the most frequently used platform for this purpose (Paolacci & Chandler, 2014).

Data collection using MTurk allows researchers to access a large and diverse pool of data with the benefits of high speed data collection and at a relatively low cost (Aquinis et al., 2021). Researchers argue that data collected using MTurk exhibits high reliability, and this method is, therefore, considered a valid data collection technique (Casler et al., 2013; Johnson & Borden, 2012). A study using a meta-analytic approach showed that data from MTurk provide effect size estimates similar to the conventional data and achieve the internal and external validity while arguing that the sample source is able to manage the research questions (Walter et al., 2019).

3.3. Data Analysis

The occurrences of best and worst selections for each attribute were tabulated into Best/Most and Worst/Least frequencies from each set. In the 26 choice sets, each attribute can be selected either as Best/Most item six times or as Worst/Least item six times. The Best/Worst (BW) score is regarded as the total worst score subtracted from its total best score, ranging from +6 to –6. The next estimated value is the Average BW (ABW) score calculated by dividing the total BW scores by the number of respondents and the frequency of replication. The rankings of attributes are generated according to the BW and ABW scores in the tables. The formula of ABW is as follows:

ABW score = [Count Best (Most) - Count Worst (Least)]/a ×n

Count Best (Most) = the total number of attributes chosen as the most important

Count Worst (Least) = the total number of attributes chosen as the least important

a = frequency of replication of each attribute

n = the number of total respondents

In order to notify choice probability of each attribute, the ratio scores of attributes for relative importance can be calculated by setting the most important attribute among listed attributes as the benchmark of 100% (Auger et al., 2007; Flynn et al., 2007; Lee et al., 2008; Marley & Louviere, 2005). To avoid dividing by zero, 0.5 is added to the worst score, and the value of relative importance interprets the percentage that an attribute is likely chosen best as the most important (Cohen, 2009). The formula for relative importance is shown below.

RI = SQRT [Count Best (Most)/(Count Worst (Least)+0.5)]

4. Results

4.1. Respondents' Profile

As shown in Table 2 of respondent profiles, about 60% of respondents are male whereas 40% are female; 46% of them hold a bachelor's degree, and 73.2% are reported as White Caucasian. The most prominent age range is between 21 and 30 (51.7%), followed by an age range between 31 and 40 (33.8%). Respondents' income ranges are relatively evenly distributed from "US$ 40,000–59,999" to "US$ 100,000 or more" with around 31% of respondents reporting an income of "less than US$40,000". The average number of respondents' international trips per year is about 1.18, and the average room rate for P2P accommodation rentals per night is US$143.33.

[Table 2]

4.2. Host Attributes

Shown in Figure 1, the bar graph indicates the importance levels of host attributes identified by BWS while the line graph displays the mean value of each attribute by Likert scale. Regarding the result of BWS, the first seven bars filled in black are identified to be "most important" host attributes, whereas the six bars in gray are "least important" ones. The order of the attributes indicates the importance levels with the best attribute being "overall review scores" and the worst being "host age." Results by Likert scale indicate that the most important host attribute is "overall review scores" (4.37), followed by "host identity verified" (4.26), "number of reviews" (4.19), "number of photos" (4.15), "response time" (3.94), "response rate" (3.94), and "superhost status" (3.17). The other six attributes, the average value of which is below 3.0, are namely "full-time vs. part-time host" (2.97), "languages" (2.96), "multi-listing vs. single-listing host" (2.90), "host's personal picture" (2.67), "host age" (2.31), and "host gender" (2.26).

[Figure 1]

Although both Likert scale and BWS approach identify the seven most important attributes, the importance levels and rankings vary. For example, "number of reviews"

(ABW: 0.505) and “host identity verified” (ABW: 0.502) are ranked as 2nd and 3rd by the results of BWS, whereas the results by Likert scale show that the mean of “number of reviews” (Mean: 4.19) is lower than “host identity verified” (Mean: 4.26). Also, “superhost status” (ABW: 0.107) and “response rate” (ABW: 0.101) are ranked as 6th and 7th respectively, using BWS, but the results by Likert scale show that the mean value of “response rate” (3.94) is higher than “superhost status” (3.17). Another interesting difference between the Likert scale and BWS is the results of “response time” and “response rate”. The results of the Likert scale show the same mean values (3.94) for “response time” and response rate”, but “response time” is ranked 5th and “response rate” ranked 7th by BWS.

4.3. Accommodation Attributes

Displayed in Figure 2, the means of all accommodation attributes by Likert scale are greater than 3.0 (neutral). The most important attribute is “price” (4.46), followed by “location” (4.37), “accommodation type” (4.12), “amenities” (4.07), “number of bedrooms” (3.87), “house rules” (3.87), “cleaning fee” (3.76), “number of bathrooms” (3.75), “check in/out time” (3.71), “maximum number of guests” (3.70), “cancellation policy” (3.70), “minimum length of stay” (3.66), and “instant bookable” (3.56). The results by Likert scale show that all accommodation attributes are perceived as important factors. Unlike the previous results, it was shown by BWS that only four important accommodation attributes were highlighted as salient factors, namely “price”, “location”, “accommodation type” and “amenities”, whereas nine attributes are relatively unimportant. This result highlights the strength of BWS, as BWS can distinguish the relative importance levels among attributes while most scaling methods do not indicate the relative importance derived from the maximum utilities of trade-off options. In addition, the unimportant attributes identified by BWS are ranked differently from the results by Likert scale. For example, “cleaning fee” is ranked 12th by BWS, but 6th by Likert Scale.

[Figure 2]

4.4. Gender Differences on Host Attributes

Gender differences in perceived importance of Airbnb accommodation attributes were explored to compare the results by two scaling approaches. First, the means of Likert scales between male and female are compared using a t-test. As shown in Table 3, several significant differences between male and female users are found in host attributes. For example, “host identity verified” is the most important host attribute for males, it was ranked as 4th for females. For female users, “overall review scores” is the most important host attribute. Further, gender differences are found for “overall review scores”, “number of reviews”, “number of photos”, “response rate”, and “response time”. In general, the mean values from female respondents were higher than those from male respondents.

[Table 3]

The results of gender differences by BWS are also displayed in Table 4. Seven host attributes are identified as important for males and females and six attributes are recognized as unimportant. The most important attribute for both males and females is “overall review scores”, however, the other six attributes are listed in different rankings. For example, “host identity verified” is more important for male users and ranked 2nd, while it is ranked 3rd by female users. Similarly, “number of reviews” is more important for females and ranked 2nd, while it is ranked 3rd by males.

[Table 4]

4.5. Gender Differences in Accommodation Attributes

Table 5 exhibits the gender differences in the perceived importance of accommodation attributes based on the results of t-tests. Unlike host attributes (with the exception of a single attribute, “number of bathrooms”), significant gender differences in perceived importance were found among 12 accommodation attributes. In general, the average scores in the female group are higher than those in the male group.

[Table 5]

The results by BWS revealed several interesting findings. As shown in Table 6, only four attributes are important for both groups (i.e., “price” “location” “accommodation type” and “amenities”), “number of bedrooms” is identified as important only for the male group while “house rules” is important only for the female group.

[Table 6]

5. Discussion

The purpose of this study is to compare results generated by using two distinct scales (i.e., BWS and Likert scales) within a single study in the P2P Accommodation sharing context. This study provides fruitful methodological implications and extends the literature with several findings concerning both the use of BWS and Likert scales considering P2P accommodation attributes. Both scales can be useful based on specific research needs. The key point of this study was not to verify which method is superior to the other, but rather how they differ from each other.

5.1. Implications

While this study focuses particularly on cases where researchers wish to identify the relative importance of the items in which they are interested, there may indeed be instances where both Likert scales and BWS may be used within the same survey, with Likert scales being used to rate the importance of each item individually, followed by BWS being used to rate the relative importance of those items that were rated highly but similarly using Likert scales.

Much research has focused on identifying hotel selection factors based on accommodation attributes sought by guests and the importance of those attributes was usually measured using Likert-type scales, asking respondents to rate the importance of certain attributes to the consumer choices (e.g., a scale in which 1 is labeled as 'extremely unimportant' and 5 is labeled as 'extremely important'). Although respondents can easily understand the Likert scale as the most common approach for survey collection, the reliability of the results can be questionable. As each item must be rated individually, this can result in large numbers of items that must be rated and, thus, lengthy surveys. This, in turn, can lead to cognitive overload resulting in respondent fatigue and "straightlining" of answers (i.e., respondents giving the same rating to every item). Miller (1956) showed that the average human mind is capable of distinguishing about seven different items. Further, in the case of many items being given the same or very similar ratings in terms of importance, it becomes impossible to determine their relative importance.

When it comes to booking accommodation, consumers have greater access to information about the properties and services than ever before. The increasing number of options available and the increasing amount of decision-relevant information leads to consumer confusion, especially if the information is too similar, too complex, or too much (Turnbull et al., 2000). When consumers plan their trip, they may become overwhelmed by too many choices, similar services and information, and the increasing complexity of options. Therefore, it is important for service providers to highlight the attributes that are most valued by their target market. Meanwhile, researchers in tourism want to identify the key attributes, because not all attributes are equally important in determining consumer choice. While a small number of items can be evaluated using paired comparisons, it is not workable for respondents to evaluate all possible items in survey settings when the number of objects to compare grows. The BWS method can readily handle a large number of items by reducing the number of choices in a set by adopting a BIBD design.

5.2. Methodological Contributions

Scale development and validation of measures have continued to be challenging activities. Social scientists have developed several valid measures for an array of abstract concepts, particularly in the tourism and leisure fields. First and foremost, this study touched upon a theory-based scaling method, BWS, and compared it with Likert scales to uncover the methodological implications of BWS. As the name of the scale indicates, BWS is unrivaled in the study objective to verify the importance levels among multiple items.

This study shows that BWS results help us to validate priorities from the customer perspective by achieving better discrimination among existing and/or new attributes. The literature is, thereby, extended by building on the work of Baumgartner and Steenkamp (2001) and Lee et al. (2007). The cognitive burden of BWS tends to be light, provided that the total number of annotations within BWS is limited.

Likert scales are still useful when researchers want to collect data from people familiar with Likert scales, as it is the most widely adopted scale. Early career researchers, including graduate students, may also like to use the Likert scale approach as it is simple to administer and score, and the responses are easily quantifiable. The Likert scale approach is useful for measuring each of the estimated values on importance so that further comparisons such as gender disparities on perceived importance levels using t-tests can be conducted. Likert scales also allow for indifference choices (i.e., two items being given the same ranking), which BWS does not, and allows for individual items to be ranked according to different attributes (e.g., most important/least important; most valuable/least valuable; too few/too many;

realistic/unrealistic; etc.).

On the other hand, the BWS method may offer advantages to cross-cultural researchers in terms of cross-cultural equivalence. While Likert's approach includes multiple verbal scale terms, the BWS approach has only two verbal scale terms (most important and least important). Therefore, the problems associated with lexical equivalence can be reduced as it is easier to find equivalent terms for "most" and "least" in different languages. However, the BWS approach would be relatively limited if further statistical analyses, such as a testing cause-effect relationship, are required. The detailed strengths and weaknesses of the Likert scale and BWS are summarized in Appendix 2.

Regarding tourism and hospitality management research, the BWS method can be applied to identify the key attributes of broad concepts or the core items of complex systems on which to focus. The BWS method also helps to translate academic research findings into practical applications. For example, sustainable tourism has been a popular research topic, but it is a comprehensive concept that involves natural environments as well as the economic and social impact on local communities. Although many research efforts have been devoted to this topic - and embedding sustainability into tourism business has become common practice - industry practitioners are still keen to know how to prioritize practices and which practices to adopt first in order to build a competitive advantage.

The BWS method can be useful in responding to such inquiries. In addition, BWS questionnaires can be used for general visitor surveys, as BWS has been found to be relatively easy for respondents to understand and answer. Moreover, the cognitive burden is relatively lighter than alternative scales such as Likert. In addition, as BWS is powerful with a smaller sample size and less affected by external factors, hospitality and tourism scholars may consider using the BWS method when dealing with small sample sizes. Importantly, this straightforward method is justified by the random utility theory (Flynn & Marley, 2014), enabling hospitality and tourism researchers to emphasize strong theoretical foundations in applying the method to relative choice probability among attributes in order to evaluate the significance level in a more accurate way than any other conventional scales.

6. Limitations and Future Research

As this study compares only two scaling approaches, it is strongly recommended that future researchers apply other research methods such as conjoint analysis or discrete choice modeling approach to estimate the value of each attribute. For example, Huertas-Garcia et al (2014) examined the significance of hotel-related attributes in affecting guests' decision-making using conjoint analysis. Masiero et al (2015) used a stated choice experiment and discrete choice modeling method to obtain hotel guests' willingness to pay for a specific set of hotel room attributes. Further studies may apply different scaling approaches in various contexts to generate empirical findings that have contributed to a deeper understanding of the rationale behind consumer choices. Several further questions should be addressed in future research, including how both Likert scales and BWS may be used within a single survey. Flynn et al. (2007) mentioned that there is no general guideline for defining a sufficient sample size for a BWS approach. Future research may provide further guidelines to calculate sample size for a BWS study.

REFERENCES

Aquinis, H., Villamor, I., & Ramani, R. S. 2021. “MTurk research: Review and recommendations.” Journal of Management 47(4): 823-837.

Arceneaux, K. 2012. “Cognitive biases and the strength of political arguments.” American Journal of Political Science 56(2): 271-85.

Auger, P., Devinney, T. M., and Louviere, J. J. 2007. “Using best–worst scaling methodology to investigate consumer ethical beliefs across countries.” Journal of Business Ethics 70(3): 299-326.

Baumgartner, H., and Steenkamp, J. B. E. 2001. “Response styles in marketing research: A cross-national investigation.” Journal of Marketing Research 38(2): 143-56.

Berinsky, A. J., Margolis, M. F., and Sances, M. W. 2014. “Separating the shirkers from the sorkers? Making sure respondents pay attention on self-administered surveys.” American Journal of Political Science 58(3): 739-53.

Bertram, D. 2007. “Likert scales.” Retrieved November 2(10), 1-10.

Burton, N., Burton, M., Fisher, C., Peña, P. G., Rhodes, G., and Ewing, L. (2021). “Beyond Likert ratings: Improving the robustness of developmental research measurement using best–worst scaling. ” Behavior Research Methods 1-7.

Burke, P., Schuck, S., Aubusson, P., Buchanan, J., Louviere, J., and Prescott, A. 2013. “Why do early career teachers choose to remain in the profession? The use of best–worst scaling to quantify key factors.” International Journal of Educational Research 62; 259-68.

Casler, K., Bickel, L., and Hackett, E. 2013. “Separate but equal? A comparison of participants and data gathered via Amazon’s MTurk, social media, and face-to-face behavioral testing.” Computers in Human Behavior 29(6): 2156-60.

Chattopadhyay, M., and Mitra, S. K. 2020. “What Airbnb host listings influence peer-to-peer tourist accommodation price?” Journal of Hospitality & Tourism Research 44(4): 597-623.

Chrzan, K., and Skrapits, M. 1996. “Best−Worst Conjoint Analysis: An Empirical Comparison with a Full Profile Choice-Based Conjoint Experiment.” In proceeding of the INFORMS Marketing Science Conference. Gainesville, FL.

Chu, R. K., and Choi, T. 2000. “An importance-performance analysis of hotel selection factors in the Hong Kong hotel industry: a comparison of business and leisure travelers.” Tourism Management 21(4): 363-77.

Cohen, E. 2009. “Applying best−worst scaling to wine marketing.” International Journal of Wine Business Research 21(1): 8-23.

Cohen, S. H. 2003. “Maximum Difference Scaling: Improved measures of importance and preference for segmentation.” Sawtooth Software Conference Proceedings, Sequim, WA. https://www.sawtoothsoftware.com/download/techpap/maxdiff.pdf (accessed March 1, 2020).

Cohen, S. H., and Markowitz, P. 2002. “Renewing market segmentation: Some new tools to correct old problems.” In Proceeding of the ESOMAR 2002 Congress. Amsterdam, Netherlands

Cohen, S. H., and Neira, L. 2003. “Measuring preferences for product benefits across countries: Overcoming scale usage bias with maximum difference scaling.” ESOMAR 2003 Latin America Conference Proceedings, Amsterdam, Netherlands

Crask, M. R., and Fox, R. J. 1987. “An exploration of the interval properties of 3 commonly used marketing-research scales- a magnitude estimation approach.” Journal of the Market Research Society 29(3): 317-39.

Dittrich, R., Francis, B., Hatzinger, R. and Katzenbeisser, W. 2007. “A Paired Comparison

Approach for the Analysis of Sets of Likert-scale Responses." Statistical Modelling 7(1): 3-28.

Edelman, B. G., and Luca, M. 2014. "Digital discrimination: The case of Airbnb.com." Harvard Business School NOM Unit Working Paper, (14-54).

Ert, E., Fleischer, A., and Magen, N. 2016. "Trust and reputation in the sharing economy: The role of personal photos in Airbnb." Tourism Management 55, 62-73.

Finn, A., and Louviere, J. J. 1992. "Determining the appropriate response to evidence of public concern: the case of food safety." Journal of Public Policy and Marketing 12-25.

Flynn, T. N., Louviere, J. J., Peters, T. J., and 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): 171-89.

Flynn, T. N., and Marley, A. A. 2014. "Best-worst scaling: theory and methods." In Handbook of choice modelling. Edward Elgar Publishing. pp.1-29

Garrido, L. E., Abad, J. J., and Ponsoda, V. 2013. "A new look at Horn's parallel analysis with ordinal variables." Psychological Methods 18(4): 454-74.

Gibson, C., E. Ostrom, and T.-K. Ahn. 2000. "The concept of scale and the human dimensions of global change: a survey." Ecological Economics 32: 217-239.

Goodman, S., Lockshin, L., and Cohen, E. 2005. "Best worst scaling: a simple method to determine drinks and wine style preferences." In Proceedings of the 2nd Annual International Wine Marketing Symposium, Sonoma State University, Rohnert Park, CA.

Harwell, M. R., and Gatti, G. G. 2001. "Rescaling ordinal data to interval data in educational research." Review of Educational Research 71(1): 105-131.

Healy, A., and Lenz, G. S. 2014. "Substituting the end for the whole: Why voters respond primarily to the election-year economy." American Journal of Political Science 58(1): 31-47.

Huertas-Garcia, R., Laguna García, M., and Consolación, C. 2014. "Conjoint analysis of tourist choice of hotel attributes presented in travel agent brochures." International Journal of Tourism Research 16: 65-75.

Hui, C. H., and Triandis, H. C. 1989. "Effects of culture and response format on extreme response style." Journal of Cross-Cultural Psychology 20: 296-309.

Jaeger, S. R., and Cardello, A. V. 2009. "Direct and indirect hedonic scaling methods: A comparison of the labeled affective magnitude (LAM) scale and best–worst scaling." Food Quality and Preference 20(3): 249-58.

Joshi, A., Kale, S., Chandel, S., and Pal, D. K. 2015. "Likert scale: Explored and explained." British Journal of Applied Science & Technology 7(4); 396-403.

Johnson, D. R., and Borden, L. A. 2012. "Participants at your fingertips using Amazon's Mechanical Turk to increase student–faculty collaborative research." Teaching of Psychology 39(4): 245-51.

Kim, B., Kim, S., King, B. and Heo, C. Y. 2019. "Luxurious or economical? An identification of tourists' preferred hotel attributes using best–worst scaling (BWS)." Journal of Vacation Marketing 25(2): 162-75.

Kiritchenko, S., and Mohammad, S. M. 2017. "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation." In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, Canada. pp.1-7.

Lagerkvist, C. J. 2013. "Consumer preferences for food labelling attributes: Comparing direct ranking and best–worst scaling for measurement of attribute importance, preference intensity and attribute dominance." Food Quality and Preference 29(2): 77-88.

Lee, J. A., Soutar, G., and Louviere, J. J. 2007. “Measuring values using best−worst scaling: the LOV example.” Psychology and Marketing 24: 1043-58.

Lee, J. A., Soutar, G., and Louviere, J. 2008. “The best–worst scaling approach: An alternative to Schwartz's values survey.” Journal of Personality Assessment 90(4): 335-47.

Li, Q. 2013. “A novel Likert scale based on fuzzy sets theory.” Expert Systems with Applications 40(5): 1609-18.

Liang, L. J., Choi, H. C., and Joppe, M. 2018. “Understanding repurchase intention of Airbnb consumers: perceived authenticity, electronic word-of-mouth, and price sensitivity.” Journal of Travel and Tourism Marketing 35(1): 73-89.

Likert, R. 1932. “A technique for the measurement of attitudes.” Archives of psychology 22(140): 55.

Lockshin, L., Cohen, E., and Zhou, X. 2011. “What influences five-star Beijing restaurants in making wine lists?” Journal of Wine Research 22(3): 227-43.

Lockshin, L., Corsi, A. M., Cohen, J., Lee, R., and Williamson, P. 2017. “West versus East: Measuring the development of Chinese wine preferences.” Food Quality and Preference 56: 256-65.

Louviere, J. J., and Islam, T. 2008. “A comparison of importance weights and willingness-to pay measures derived from choice-based conjoint, constant sum scales and best−worst scaling.” Journal of Business Research 61(9): 903-11.

Louviere, J. J., Swait, J., and Anderson, D. 1995. Best−worst Conjoint: A new preference elicitation method to simultaneously identify overall attribute importance and attribute level partworths. University of Sydney. https://pdfs.semanticscholar.org/9af2/3b0b672a8d623489072bc9b06e19fa7885f3.pdf (accessed March 1, 2020).

Louviere, J. J., and 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): 350-67.

Lusk, J. L., and Briggeman, B. C. 2009. “Food values.” American Journal of Agricultural Economics 91(1): 184-96.

Ma, X., Hancock, J. T., Mingjie, K. L., and Naaman, M. 2017. “Self-disclosure and perceived trustworthiness of Airbnb host profiles.” Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1–13): New York.

Masiero, L., Heo, C. Y., and Pan. B. 2015. “Determining guests’ willingness to pay for hotel room attributes with a discrete choice model.” International Journal of Hospitality Management 49: 117-24.

Marley, A. A., and Louviere, J. J. 2005. “Some probabilistic models of best, worst, and best worst choices.” Journal of Mathematical Psychology 49(6): 464-80.

Miller, G. A. 1956. “The magical number seven, plus or minus two: some limits on our capacity for processing information.” Psychological Review, 63(2); 81:97.

Mueller, S., Francis, I. L., and Lockshin, L. 2009. “Comparison of best–worst and hedonic scaling for the measurement of consumer wine preferences.” Australian Journal of Grape and Wine Research 15(3): 205-15.

Nunes, F., Madureira, T., Oliveira, J. V., and Madureira, H. 2016. “The consumer trail: Applying best-worst scaling to classical wine attributes.” Wine Economics and Policy 5(2); 78-86.

Paolacci, G., and Chandler, J. 2014. “Inside the Turk: Understanding Mechanical Turk as a Participant Pool.” Current Directions in Psychological Science 23(3): 184-8.

Potoglou, D., Burge, P., Flynn, T., Netten, A., Malley, J., Forder, J., and 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.

Qu, H., Ryan, B., and Chu, R. 2000. “The importance of hotel attributes in contributing to travelers' satisfaction in the Hong Kong hotel industry.” Journal of Quality Assurance in Hospitality and Tourism 1(3): 65-83.

Scarpa, R., Notaro, S., Louviere, J., and 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): 813-28.

Thurstone, L.L. 1927. “A law comparative judgment.” Psychological Review 34: 273-86.

Turnbull, P. W, Leek, S, and Ying, G. 2000. “Customer confusion: The mobile phone market.” Journal of Marketing Management 16: 143-63.

Walter, S. L., Seibert, S. E., & Goering, D., & O’Boyle, E. H. 2019. “A tale of two sample sources: Do results from online panel data and conventional data converge?” Journal of Business and Psychology 34: 425-452.

Wang, D., and Nicolau, J. L. 2017. “Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com.” International Journal of Hospitality Management 62: 120-31.

Weller, S. C., and Romney, A. K. 1988. “Systematic data collection (Vol. 10).” Sage publications. Inc.: Washington, U.S.A.

Yasmin, F., Ahmed, R., and Akhtar, M. 2015. “Construction of Balanced Incomplete Block Designs Using Cyclic Shifts.” Communications in Statistics-Simulation and Computation 44(2): 525-32.

Table 1. Host and accommodation attributes of P2P accommodations

	Host attributes	Accommodation attributes
1	Superhost status	Accommodation type (house, apt etc.)
2	Host listings count	Maximum number of guests
3	Host’s profile picture	Number of bathrooms
4	Host identity verified	Number of bedrooms
5	Number of reviews	House rules
6	Review scores for overall rating	Amenities
7	Number of photos	Location
8	Response time	Number of beds
9	Response rate	Price
10	Languages	Instant bookable
11	Host gender	Cancellation policy
12	Host age	Minimum length of stay
13	Host race	Cleaning fee

Table 2. Respondent profile (N=302)

		%			%
Gender	Male	59.6	Age	20 or less	2.0
	Female	40.4		21-30	51.7
Education	High school or less	10.9		31-40	33.8
	College student	12.6		41-50	5.6
	Associates degree	17.2		51-60	5.3
	Bachelor’ Degree	46		61 or more	1.7
	Master’s degree	10.6	Income	Less than US$ 40,000	30.8
	Doctoral degree	2.6		US$ 40,000–59,999	21.2
Ethnicity	Caucasian	73.2		US$ 60,000–79,999	19.5
	African-American	8.6		US$ 80,000–99,999	12.3
	Hispanic	5.3		US$ 100,000 or more	16.2
	Asian	11.3	Nationality	American	98.7
	Other	1.7		Other	1.3

	Mean	Std. Deviation
Frequency of international travel per year	1.18	1.395
Average room rate spent at a hotel per night	143.33	114.338

Table 3. Host attributes between male and female datasets using Likert scale

	Male (n=180)		Female (n=122)		t-test	
	Mean	Std	Mean	Std	t	sig
Host identity verified	4.22 (1)	1.048	4.33 (4)	1.032	-.910	.363
Overall review scores	4.21 (2)	1.052	4.61 (1)	.807	-3.766	.000***
Number of reviews	4.03 (3)	1.030	4.41 (2)	.888	-3.292	.001**
Number of photos	4.01 (4)	1.028	4.36 (3)	.824	-3.185	.002**
Response rate	3.81 (5)	1.040	4.14 (6)	.753	-3.180	.002**
Response time	3.79 (6)	1.062	4.16 (5)	.843	-3.335	.001**
Superhost status	3.13 (7)	1.121	3.23 (7)	1.059	-.748	.455
Full-time vs. Part-time host	3.02 (8)	1.088	2.89 (9)	1.225	.959	.338
Multi-listing vs. Single-listing host	2.99 (9)	1.014	2.77 (10)	1.134	1.713	.088
Languages	2.91 (10)	1.240	3.03 (8)	1.233	-.877	.381
Host's personal picture	2.66 (11)	1.233	2.67 (11)	1.295	-.075	.941
Host age	2.34 (12)	1.197	2.26 (13)	1.205	.584	.560
Host gender	2.19 (13)	1.149	2.36 (12)	1.273	-1.181	.239
Mean average	3.33		3.48			

* Note: sorted out according to rankings of results of male dataset, the brackets in the column of means of the female dataset indicate the descending order according to means * $p < .05$, ** $p < .01$, ***$p < .001$

Table 4. Host attributes between male and female datasets using BWS

Male (n=180)						Female (n=122)					
	Best	Worst	B-W	ABW	Relative importance		Best	Worst	B-W	ABW	Relative importance
Overall review scores	811	70	741	0.686	100.0	Overall review scores	534	59	475	0.649	100.0
Host identity verified	666	99	567	0.525	76.3	Number of reviews	447	84	363	0.496	76.8
Number of reviews	675	123	552	0.511	68.9	Host identity verified	428	85	343	0.469	74.7
Number of photos	433	189	244	0.226	44.6	Response time	327	184	143	0.195	44.4
Response time	446	288	158	0.146	36.7	Number of photos	275	148	127	0.173	45.4
Superhost status	451	330	121	0.112	34.4	Response rate	244	159	85	0.116	41.3
Response rate	341	243	98	0.091	34.9	Superhost status	304	231	73	0.100	38.3
Full-time vs. Part-time host	249	401	-152	-0.141	23.2	Full-time vs. Part-time host	181	271	-90	-0.123	27.3
Multi-listing vs. Single-listing host	233	462	-229	-0.212	20.9	Multi-listing vs. Single-listing host	158	315	-157	-0.214	23.6
Host's personal picture	126	484	-358	-0.331	15.0	Languages	78	334	-256	-0.350	16.1
Languages	102	503	-401	-0.371	13.3	Host's personal picture	76	351	-275	-0.376	15.5
Host gender	88	729	-641	-0.594	10.2	Host gender	65	480	-415	-0.567	12.3
Host age	59	759	-700	-0.648	8.2	Host age	55	471	-416	-0.568	11.4

* Note: the areas shadowed in gray in the table indicate the important host attributes identified by BWS.

Table 5. Accommodation attributes between male and female datasets using Likert scale

	Male (n=180)		Female (n=122)		t-test	
	Mean	Std	Mean	Std	t	sig
Price	4.31 (1)	1.068	4.70 (1)	.679	-3.889	.000***
Location	4.21 (2)	1.076	4.61 (2)	.698	-4.007	.000***
Amenities	3.95 (3)	1.048	4.24 (4)	.772	-2.593	.010*
Accommodation type (house, apt etc.)	3.91 (4)	1.122	4.43 (3)	.760	-4.882	.000***
House rules	3.77 (5)	1.031	4.03 (6)	.852	-2.444	.015*
Number of bedrooms	3.76 (6)	1.034	4.03 (6)	1.004	-2.314	.021*
Number of bathrooms	3.70 (7)	1.002	3.82 (11)	1.099	-.979	.328
Check in/out time	3.59 (8)	1.123	3.89 (9)	.907	-2.597	.010*
Maximum number of guests	3.59 (9)	1.092	3.86 (10)	1.093	-2.121	.035*
Cleaning fee	3.57 (10)	1.078	4.05 (5)	.908	-4.150	.000***
Cancellation policy	3.56 (11)	1.084	3.91 (8)	.900	-3.039	.003**
Instant bookable	3.43 (12)	1.078	3.75 (12)	1.031	-2.581	.010*
Minimum length of stay	3.41 (13)	1.102	4.02 (7)	.975	-5.135	.000***
Mean average	3.75		4.10			

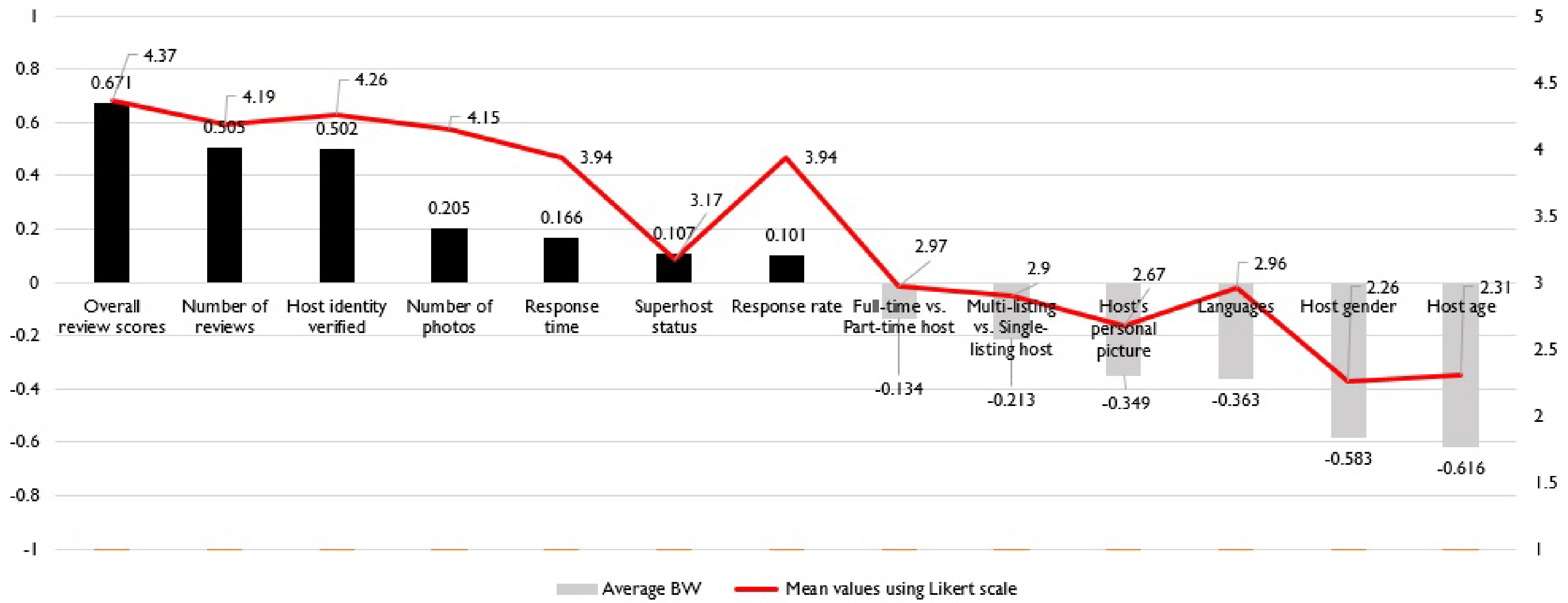
* Note: sorted out according to rankings of results of male dataset, the brackets in the column of means of the female dataset indicate the descending order according to means * $p < .05$, ** $p < .01$, ***$p < .001$

Table 6. Accommodation attributes between male and female datasets using BWS

Male (n=180)						Female (n=122)					
	Best	Worst	B-W	ABW	Relative importance		Best	Worst	B-W	ABW	Relative importance
Price	746	108	638	0.591	100.0	Price	518	65	453	0.619	100.0
Location	643	134	509	0.471	83.4	Location	374	105	269	0.367	67.0
Accommodation type (house, apt etc.)	555	232	323	0.299	58.9	Accommodation type (house, apt etc.)	371	160	211	0.288	54.1
Amenities	457	204	253	0.234	57.0	Amenities	306	125	181	0.247	55.5
Number of bedrooms	318	315	3	0.003	38.3	House rules	239	226	13	0.018	36.5
House rules	292	369	-77	-0.071	33.9	Number of bedrooms	212	241	-29	-0.040	33.3
Number of bathrooms	296	460	-164	-0.152	30.6	Maximum number of guests	211	321	-110	-0.150	28.8
Check in/out time	259	446	-187	-0.13	29.0	Check in/out time	188	300	-112	-0.153	28.1
Maximum number of guests	283	479	-196	-0.181	29.3	Number of bathrooms	180	343	-163	-0.223	25.7
Cancellation policy	210	415	-205	-0.190	27.1	Cancellation policy	129	292	-163	-0.223	23.6
Minimum length of stay	218	474	-256	-0.237	25.8	Instant bookable	152	328	-176	-0.240	24.2
Cleaning fee	192	497	-305	-0.282	23.7	Minimum length of stay	160	345	-185	-0.253	24.2
Instant bookable	211	547	-336	-0.311	23.7	Cleaning fee	132	321	-189	-0.258	22.8

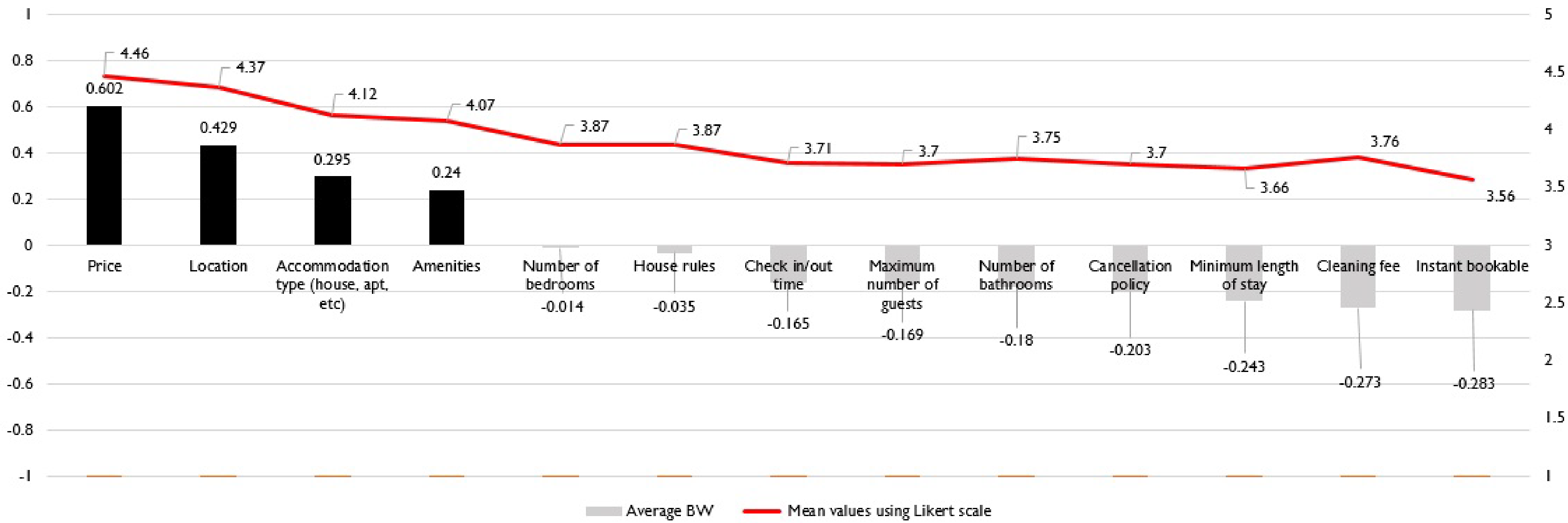
* Note: the areas shadowed in gray in the table indicate the important host attributes identified by BWS.

Figure 1. Importance levels of host attributes identified using BWS and Likert scale



* Note: n=302, the numbers for BWS results indicate Average BW (between -1 and 1); grand mean by Likert scale is 3.39.

Figure 2. Importance levels of accommodation attributes identified using BWS and Likert scale



* Note: n=302, the numbers for BWS results indicate Average BW (between -1 and 1); grand mean by Likert scale is 3.89.

Appendix 1. Summary of previous studies on BWS

Author(s)	Context	Findings
Finn & Louviere (1992)	Food safety	- Measures public concerns about food safety
Goodman et al. (2005)	Wine choice and wine style preference	- Identifies the most important attributes for wine selection and wine style preferences in two different countries - Discovers different patterns of choice between groups
Burke, et al., (2013)	Career choice	- Quantifies the relative importance of these factors that influence a teacher's decision to remain in the profession
Flynn, et al. (2007)	Quality-of-life	- Demonstrates how richer insights can be drawn by the use of BWS using a quality-of-life pilot study
Cohen (2009)	Wine choice	- Exhibits the BWS method by an empirical example of wine choice
Jaeger & Cardello (2009)	Food preference	- Compares the labeled affective magnitude (LAM) scale of liking and BWS to identify the acceptance levels of seven fruit juices and preference
Mueller, et al. (2009)	wine preferences	- Compares best–worst and hedonic scaling for consumer wine preferences - BWS has a higher discriminative ability for different products in non-sensory selections
Scarpa, et al. (2011)	Tourism benefit	- Estimates benefits of tourism in alpine grazing commons
Potoglou, et al. (2011)	Social data	- Presents empirical findings from the comparison between discrete choice experiments and profile-based best-worst scaling
Lockshin, et al. (2011)	Wine choice	- Focuses on wine preferences in making wine lists in five-star Chinese restaurants
Lagerkvist (2013)	Beef labeling	- Compares attributes of labeling of beef using BWS and standard direct ranking. - BWS showed more accurate individual choice predictions and consistent dominance ordering on attribute importance levels.
Nunes, et al. (2016)	Wine choice	- Finds extrinsic attributes that influence wine purchase choices in a retail store
Kim, et al. (2019)	Hotel selection	- Identifies hotel selection attributes between luxury and economy hotel customer segments using BWS

Appendix 2. Comparison between BWS and Likert scale

	BWS proposed by Finn and Louviere (1992)	Likert scale developed by Likert (1932)
Strength	- BWS questionnaires are relatively easy for respondents to understand and answer. - Cognitive burden for respondents is relatively light. - All attribute levels are on the same scale. - The relative values associated with each of a list of objects can be measured. - Preference structures can be determined more precisely with a smaller sample size. - The priorities among the items in the list can be validated from a given respondent's perspective - Fewer and weaker assumptions about human decision-making affected by external factors (e.g., culture, age, etc.)	- Respondents are familiar as it is the most widely adopted scale. - Likert scales are simple to administer and score. - The responses are easily quantifiable and used for computation of some mathematical analyses (e.g., group comparison, causality testing)
Weakness	- Specific study design for data collection (e.g., BIBD) is required. - As partial rankings of attributes based on sequential choices, the first response can have an influence on that of the second question. - Indifference choices are not allowed (e.g., 2 equally important attributes).	- Likert scale is uni-dimensional and only gives 5-7 options of choice. - As the space between each choice cannot possibly be the same, it fails to measure the true responses. - The responses can be on a neutral point when participants do not have a specific opinion. - The usage of Likert scale can be cognitively demanding for participants. - Respondents' answers can be influenced by previous questions. - Statistical power, such as large sample size, trial numbers, should be fulfilled to prove the robustness of conclusions. - Discrimination among attributes can be identified by only using the individual number without comparing relative importance.