Equalizing bias in eliciting attribute weights in multiattribute decision-making experimental research

Rezaei, Jafar; Arab, Alireza; Mehregan, Mohammadreza

DOI
10.1002/bdm.2262

Publication date
2021

Document Version
Final published version

Published in
Journal of Behavioral Decision Making

Received: 4 January 2021 | Revised: 1 August 2021 | Accepted: 2 August 2021

DOI: 10.1002/bdm.2262

RESEARCH ARTICLE

WILEY

Equalizing bias in eliciting attribute weights in multiattribute decision-making: experimental research

Jafar Rezaei[1] | Alireza Arab[2] | Mohammadreza Mehregan[2]

[1]Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

[2]Department of Industrial Management, Faculty of Management, University of Tehran, Tehran, Iran

Correspondence
Jafar Rezaei, Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX, Delft, The Netherlands.
Email: j.rezaei@tudelft.nl

Abstract

One of the most important steps in formulating and solving a multiattribute decision-making (MADM) problem is weighting the attributes. Most existing weighting methods are based on judgments by experts/decision-makers, which are prone to several cognitive biases, making it necessary to examine these biases in MADM weighting methods and develop debiasing strategies. This study uses experimental analysis to look at equalizing bias—one of the main cognitive biases, where decision-makers tend to assign the same weight to different attributes—in MADM methods. More specifically, we look at AHP (analytic hierarchy process), BWM (best-worst method), PA (point allocation), SMART (simple multiattribute rating technique), and Swing methods under two structuring formats, hierarchical and non-hierarchical. To empirically examine the existence of equalizing bias in these methods, we formulate several hypotheses, which are tested using a public transportation mode selection problem among 146 university students. The results indicate that AHP and BWM have less equalizing bias than SMART, Swing, and PA, and that the hierarchical problem structuring leads to a reduction in the equalizing bias in all five methods and that such a reduction significantly varies among the methods. Our findings prove some debiasing strategies suggested in existing literature, which could be used by real decision-makers (when selecting a method) as well as researchers (when developing new methods).

KEYWORDS
cognitive bias, equalizing bias, experimental analysis, multiattribute decision-making, weighting

1 | INTRODUCTION

Decision-making involves evaluating and choosing among alternative actions (Simon et al., 1987), which means that decision-makers compare these alternatives based on various aspects called 'attributes'. These attributes are of different importance for different decision-makers, which could lead to different decision-makers ending up with different best choices when considering the same set of alternatives. People usually rely on their own judgments when considering the importance of different attributes, in what is known as 'judgment of importance' (Pajala et al., 2019) and they usually use strategies, which are called 'heuristics' (Tversky & Kahneman, 1974). Generally speaking, heuristics are helpful, but their use can sometimes lead to severe errors (Bazerman & Moore, 2012) (also known as bias, or cognitive bias). For instance, anchoring and adjustment is a heuristic that involves starting from an initial value and adjusting that value to arrive at the final estimate. These adjustments are usually insufficient, because different starting points yield different estimates that are biased toward the initial values (Kahneman et al., 1982; Tversky & Kahneman, 1974). According to Bazerman and Moore (2012), becoming aware of the potential adverse impact of using heuristics makes it possible to determine when and where to use them. In other words, it provides the possibility of controlling and managing the occurrence of biases discussed above in the decision-making process.

In many real-world decision-making problems, decision-makers face multiattribute decision problems. Multiattribute decision-making (MADM) methods have been developed to deal with these kinds of problems. One of the most important steps in formulating and solving an MADM problem is attribute weighting. Researchers have developed a number of MADM weighting methods, most of which are based on judgments by experts/decision-makers to deal effectively with real-life problems that are not characterized by objective measures. Human judgments are prone to several cognitive biases, which could lead to suboptimal or even nonfeasible solutions, making it necessary to examine these biases in MADM weighting methods.

According to Marttunen et al. (2018), Rezaei (2021), and Montibeller and von Winterfeldt (2015a), there are a few researchers who have discussed cognitive biases in MADM. In one of the earliest studies in this area, Gabrielli and von Winterfeldt (1978) examined the sensitivity of the weights of the attributes to the changes of alternatives range in an experimental study, using the SMART (simple multiattribute rating technique) weighting method. The results showed that people have plausible ranges in mind when weighting attributes and are unwilling to adjust weights after relatively spurious changes occur in the set of alternatives. von Nitzsch and Weber (1993), Fischer (1995), Pöyhönen and Hämäläinen (2000), and Lin (2013) conducted similar experiments involving range insensitivity bias. Fischer et al. (1987) examined and confirmed the hypothesis that the proxy attributes gain more weight than the fundamental attributes. They proposed several debiasing strategies. Weber et al. (1988), Borcherding and von Winterfeldt (1988), Pöyhönen and Hämäläinen (1998), Pöyhönen and Hämäläinen (2000), Pöyhönen and Hämäläinen (2001), Jacobi and Hobbs (2007), and Hämäläinen and Alaja (2008) examined splitting bias, in which presenting an attribute in greater detail may increase the assigned weight. Buchanan and Corner (1997) and Rezaei (2021), in their examination of anchoring bias, had as their main finding the role the structure of the solution method played in relation to the incidence of anchoring bias. Lahtinen et al. (2020) proposed four debiasing strategies to be embedded in MADM methods to reduce the effects of framing effect, loss aversion, and status quo-type cognitive biases.

In addition to the studies listed above, some researchers, including Montibeller and von Winterfeldt (2015a), Montibeller and von Winterfeldt (2015b), Hämäläinen (2015), Montibeller and von Winterfeldt (2018), Montibeller (2018), and Marttunen et al. (2018), have conducted comprehensive reviews of this area and described all cognitive biases that can occur in each of the MADM steps, and recommended a number of debiasing strategies to mitigating the potential biases. They considered equalizing bias (the tendency of decision-makers to assign the same weight to different attributes) to be one of the main cognitive biases in weighting methods. In cases involving equalizing bias, the weights do not reflect the decision-makers (DMs) judgments that could lead to a nonsatisfactory result for decision-makers. As such, their expert opinions become less useful when applying MADM weighting methods. Despite the adverse consequences of equalizing bias, it has yet to be studied in an experimental setting and has only been mentioned in the studies mentioned above.

As such, the aim of this study is to examine equalizing bias in MADM weighting methods. The main contribution of this study is to examine equalizing bias in the eliciting of attribute weights in five MADM methods, namely, AHP (analytic hierarchy process) (Saaty, 1977), BWM (best-worst method) (Rezaei, 2015, 2016), PA (point allocation) (Doyle et al., 1997), SMART (Edwards, 1977), and Swing (von Winterfeldt & Edwards, 1986), by conducting an experimental study. We also compare these methods to demonstrate how methods with different characteristics could lead to a lower or higher level of equalizing bias, and we examine the impact of problem structuring (hierarchical vs. non-hierarchical) on the occurrence of equalizing bias. In doing so, we want to help decision-makers who are selecting an MADM weighting method by making them aware of potential bias and its effect on final attributes weight and to help researchers consider the issues raised in this study when developing new methods, allowing them to minimize the risk of this bias occurring. We also think that the findings of this study could help researchers find ways to improve the existing methods in order to make them less vulnerable to equalizing bias.

In Section 2, we discuss equalizing bias. In Section 3, the five MADM weighting methods are outlined. The research hypotheses are formulated in Section 4. An experimental analysis conducted to test the research hypotheses and check the equalizing bias of the methods discussed in Section 5. Section 6 contains the data analysis and discussion, and Section 7 contains the conclusion and future research suggestions.

2 | EQUALIZING BIAS

In the context of multiattribute weighting, equalizing bias can be defined as a tendency among decision-makers to express (about) equal judgment of importance for a set of *n* attributes, which can also be defined as the 1/*n* rule (Jacobi & Hobbs, 2007; Montibeller & von Winterfeldt, 2015b). Such a bias can also be due in part to the features of the weighting method involved. Although it is clear that there are problems where a decision-maker is truly indifferent to the importance of the attributes in question (which also results in equal weights), if we systematically encountered this phenomenon among a relatively large sample of subjects when using different weighting methods, we could draw conclusions about the source of equalizing bias in relation to the features of the procedure of the weighting method. Some studies have discussed equalizing bias in relation to MADM, most of which are included in the literature review papers presented below.

Montibeller and von Winterfeldt (2015a, 2015b, 2018) conducted a comprehensive literature review involving cognitive bias in MADM methods. They argued that equalizing bias is a relevant cognitive bias in the elicitation of attribute weights task of multiattribute analysis and proposed some debiasing strategies, including ranking of events or objectives first, then assigning ratio weights, and eliciting weights hierarchically. Marttunen et al. (2018) conducted a meta-analysis study including 61 environmental and energy cases to examine whether earlier findings regarding MADM-related cognitive biases can be found in real-world applications. In one of their investigations, they analyzed support for the equalizing bias and compared the lowest and highest weights of top-level objectives, and found no evidence of the equalizing bias. Marttunen et al. (2018) indicated that some weighting methods, for instance PA, may be more prone to equalizing bias than others. Also, in cases where there are large differences in the impact ranges over the objectives, equalizing bias can significantly distort the results. Tervonen et al. (2017) reviewed and critically assessed similarities and differences of Swing weighting and DCE (discrete choice experiments) to elicit patient benefit–risk preferences, and argued that the direct matching task involved in the Swing method makes it more prone to equalizing bias. Jacobi and Hobbs (2007) used equalizing bias as a starting point in their model for estimating and correcting objectives hierarchy induced biases in their study about quantifying and mitigating the splitting bias and other value tree-induced weighting biases. Their result showed flatter, less varied weights for non-hierarchical assessments. Stillwell et al. (1987) and Pöyhönen and Hämäläinen (1998) compared hierarchical and non-hierarchical weighting methods for eliciting multiattribute value models in experimental research, one of their findings being that hierarchical weights were steeper than non-hierarchical weights.

The literature review presented above shows that, despite the importance of equalizing bias in MADM weighting, its presence in MADM weighting methods has yet to be examined empirically.

3 | MULTIATTRIBUTE DECISION-MAKING

Multiattribute decision-making (MADM) involves evaluating a number of alternatives (options) that are characterized by a number of attributes. The evaluation can be conducted for different purposes, like selecting, ranking or sorting the alternatives. In most MADM problems, the aim is to identify and quantify the relative importance (weight) of the attributes. For a better understanding of the way theses attribute weights are used in MADM methods, here we use the example of multiattribute value theory (MAVT) (Keeney & Raiffa, 1976) as one of the most widely used methods to solve MADM problems (Weber & Borcherding, 1993). Under different conditions, additive, multiplicative or other nonadditive value functions can be used to aggregate the preferences. Suppose that we have m alternatives ($i = 1, 2, \ldots, m$), n attributes ($j = 1, 2, \ldots, n$) and $a_k = (a_{k1}, a_{k2}, \ldots, a_{kn})$ be alternative outcomes with respect to n attributes. Let $v(a_k)$ be a decision-maker's additive value function of alternative a_k, then the overall value of alternative a_k can be found as follows (Keeney & Raiffa, 1976):

$$v(a_k) = \sum_{j=1}^{n} w_j v_{kj}(a_{kj}), \tag{1}$$

where $v_{kj}(a_{kj})$ is the normalized value of a_{kj}, and w_j is the importance weight of attribute j, and $w_j > 0, \sum_{j=1}^{n} w_j = 1$. Comparing two alternatives, an alternative is preferred to another alternative if and only if its additive value (equation 1) be greater than that of the other alternative.

The additive value function can be used if the attributes are (i) mutually preferentially independent and (ii) difference independent (the definitions are from Keeney and Raiffa (1976), Currim and Sarin (1984), and Dyer and Sarin (1979)).

Definition 1. The attributes $X_1, \ldots, X_n$, are mutually preferentially independent if any subset of attributes is preferentially independent of the remaining attributes.

Definition 2. The attribute X_j is difference independent of the remaining attributes if the preference difference between two levels of X_j is not affected by the fixed levels on the other attributes.

If a weaker condition holds (mutually preferentially independence and weak difference independence), we could use a multiplicative or other nonadditive value functions (Dyer & Sarin, 1979).

Definition 3. X_j is weak difference independent of the remaining attributes if the ordering of preference differences on X_j does not depend on the fixed levels of the remaining attributes.

As stated before, calculating the importance weight of the attributes (w_j in equation 1) is one of the most important steps in MADM problems, which is the reason several MADM weighting methods have been developed in literature, including the SMART, Swing, Tradeoff (Keeney & Raiffa, 1976), AHP, ANP (analytical network process) (Saaty, 1996), and BWM, which are among the most common and widely used weight elicitation methods.

The weights obtained through these methods are based on judgments provided by people who are prone to cognitive biases. For the aim of this study, we decided to include AHP, BWM, PA, SMART, and Swing, which allowed us to appropriately cover several important features of the area under examination. AHP and BWM are representatives of pairwise comparison methods with different approaches and computational efficiency, while PA, SMART, and Swing represent scoring-based methods. SMART has a lower bound limitation, while Swing has an upper bound limitation and PA has no scoring limitation. In addition, Swing considers the range of attributes (the difference between the lowest and highest level of an attribute considering the set of all alternatives). van Ittersum et al. (2007), presented a framework proposed that what methods measure which specific dimensions of attribute importance. The three dimensions of attribute importance proposed in their framework are *salience*, *relevance*, and *determinance*. Salience refers to the attributes that come to mind more easily when someone is evaluating an alternative in cases where no attribute has been identified in advance for the evaluation, and the attributes that come to mind more easily are deemed more important. This dimension is not relevant to the aim of our study, because we do have a defined list of attributes for all the methods under study. Nevertheless, the other two dimensions are relevant. While relevance shows the importance of an attribute based on individual's 'personal values and desires', determinance shows the importance of an attribute based on 'judgment and choice'. Relevance does not consider the range of an attribute, while determinance does. Based on this framework, we can conclude that AHP, BWM, SMART, and PA infer the weights of attributes following the relevance dimension, and none of them systematically consider the range of attributes. The Swing method, on the other hand, elicits the attribute weights based on determinance dimension as it systematically takes the range of attribute into account. From a different perspective, we can also see that, while Swing systematically considers the range of attributes, other methods, of which AHP and BWM are based on pairwise comparisons and PA and SMART on the direct assignment of importance, do not. To summarize, the methods considered in this study cover a diverse spectrum of MADM methods.

Another reason for choosing this particular set of MADM methods has to do with their potentially different behavior with respect to equalizing bias (as discussed in greater detail in Section 4), which is analyzed on the basis of the comprehensive literature review we conducted in this study, based mainly on generic strategies designed to reduce equalizing bias found in literature. Two of the strategies involved are 'rank and ratio' methods and 'hierarchical weighting'. From the methods listed above, two use the rank and ratio strategy (AHP and BWM), while the other three do not, which will help us determine which type of method is better at controlling equalizing bias. In addition, we consider all the five methods in two structural formats (hierarchical and non-hierarchical), to identify potential differences in the effect of structuring format on equalizing bias (more details in Section 4).

Finally, four of the selected methods (AHP, PA, SMART, Swing) are relatively old, very popular and have been used in many studies, including cognitive bias studies, while BWM is an emerging method that has attracted considerable attention among researchers and practitioners, making them all suitable for our study.

Below, we briefly discuss the five MADM methods used in this study.

3.1 | SMART

With this method, the decision-maker first ranks the attributes in order of importance, after which the least important attribute is assigned the value of 10. Other attributes then take values greater than or equal to 10, respectively, from low to high importance. Finally, the attribute weights are calculated by normalizing the values into one by equation 2 (Pöyhönen & Hämäläinen, 2001). Suppose we have n attributes ($j = 1, 2, ..., n$), s_j is the score that decision-maker assigns to attribute j, and w_j is the importance weight of the attribute j. Then the weight of attribute j is obtained by normalizing the scores as follows.

$$w_j = \frac{s_j}{\sum_{j=1}^{n} s_j}, \forall j. \tag{2}$$

There are versions of this method that consider the range of attributes, but in this paper, similar to Rezaei (2021) and Bottomley and Doyle (2001), we intentionally use the original version of SMART, and the ranges of attributes were not described to the subjects, which we already do in Swing (see Section 3.2), providing a more diverse set of methods.

3.2 | Swing

According to this method, knowing all the alternatives, the decision-maker first identifies the best and worst level of each attribute and then is asked to consider a situation with a hypothetical alternative characterized by the worst level of all attributes, and to think about changing an attribute from its worst level to its most satisfactory level. That attribute is assigned a score of 100. The decision-maker should then find a second attribute which has the second rank of satisfaction and assign that attribute a score less than or equal to 100 (in relation to the first swing) and so on, until the worst attribute is scored. Similar to the formula used for SMART (Equation 2), the scores are normalized and then interpreted as the weights (Pöyhönen & Hämäläinen, 2001). In this paper, similar to Lin (2013) and Pöyhönen and Hämäläinen (2001), we use the original version of Swing. As becomes evident, Swing is different from SMART in two aspects: the starting point (SMART starts with the least important attribute, while Swing starts with the most important one) and the fact that SMART does not consider the range of attributes, while Swing does.

3.3 | Point allocation

In this method, the decision-maker assigns scores to attributes without following a structure like the one used in SMART or Swing. Literature suggests different ways to make these assignments. In one commonly used approach, a decision-maker is asked to divide a fixed number (say, 100) among the attributes, while another approach is simply to assign scores to attributes based on their importance, without limiting the total value (Pöyhönen & Hämäläinen, 2001). In this study, similar to Pöyhönen and Hämäläinen (2001), the second, unlimited version of this method is used. The main reason we chose the unlimited version is that we intend to make it as different as possible from the other two score assignment methods (SMART and Swing), which do have limitations (either as a lower limit: SMART with a 10, or with an upper limit: Swing with a 100). For this method, the range of attributes is not described to the subjects. The same equation as equation 2 should be then used to normalize the scores which are interpreted as weights.

3.4 | AHP

Following AHP (Saaty, 1977), the pairwise comparison matrix of the attributes is formed by the decision-maker using a pairwise comparison scale containing the numerical values between 1 to 9 (Table 1). Pairwise comparison matrix A (see equation 3) is made such that the attribute in row i, ($i = 1, 2, \ldots, n$) is compared to all attributes in columns j, ($j = 1, 2, \ldots, n$). In this matrix $a_{ji} = \frac{1}{a_{ij}}$. Assume that the problem has n attributes. For this problem, $n \times n$ pairwise comparisons are required. From this, n comparisons are $a_{ii} = 1$. The rest is $n(n-1)$, while the other half is made up of the first half's reciprocals. Finally, AHP needs $n(n-1)/2$ pairwise comparisons.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}. \tag{3}$$

The relative weights of attributes are calculated by solving equation 4, the eigenvector (W) corresponding to the largest eigenvalue (λ_{max}).

$$AW = \lambda_{max} W. \tag{4}$$

Consistency ratio (equation 5) is used to check the extent to which the decision-maker has been consistent in providing the pairwise comparisons:

$$CR = \frac{CI}{RI}, \tag{5}$$

where, random index (RI) values are given in Table 2 for different sizes of the pairwise comparison matrix (n is the number of attributes), and consistency index (CI) is calculated by equation 6:

$$CI = \frac{\lambda_{max} - n}{n-1}. \tag{6}$$

A consistency ratio not greater than 0.1 is positive evidence for informed judgment (Saaty, 1994).

TABLE 1 Scales used for AHP pairwise comparisons (Saaty, 1977)

Intensity of importance	Definition
1	Equal importance
3	Weak importance of one over another
5	Essential or strong importance
7	Demonstrated importance
9	Absolute importance
2, 4, 6, 8	Intermediate values between the two adjacent judgments

3.5 | BWM

Based on the BWM, the decision-maker first identifies a set of n attributes $\{c_1, c_2, \ldots, c_n\}$, after which best and worst attributes are determined by the decision-maker, and a pairwise comparison is made between these two attributes (best and worst) and other attributes by using a number between 1 to 9 (see Table 1). These pairwise comparisons, which are called Best-to-Others and Others-to-Worst vectors, are shown respectively as $A_B = (a_{B1}, a_{B2}, \ldots, a_{Bn})$ and $A_W = (a_{1W}, a_{2W}, \ldots, a_{nW})^T$. a_{Bj} indicates the preference of attribute B (best attribute) over attribute j and a_{jW} indicates the preference of attribute j over the worst attribute W. Then a min-max problem is formulated and solved to determine the weight of the attributes (Rezaei, 2015) as follows (equation 7).

$$\begin{aligned} & min\ max_j \left\{ \left| \frac{w_B}{w_j} - a_{Bj} \right|, \left| \frac{w_j}{w_W} - a_{jW} \right| \right\}, \\ & \text{s.t.} \quad \sum_j w_j = 1, w_j \geq 0, \forall j. \end{aligned} \tag{7}$$

This model can be solved by transferring it to the following model (8):

$$\begin{aligned} & min \xi, \\ & \text{s.t.} \left| \frac{w_B}{w_j} - a_{Bj} \right| \leq \xi, \forall j, \\ & \left| \frac{w_j}{w_W} - a_{jW} \right| \leq \xi, \forall j, \\ & \sum_j w_j = 1, w_j \geq 0, \forall j. \end{aligned} \tag{8}$$

While model (8) is a nonlinear model with possible multiple optimal solutions, there also exists a linear BWM model that provides a unique set of solutions, presented as follows (equation 9) (Rezaei, 2016).

TABLE 2 Random consistency index (Saaty, 1994)

n	0	1	2	3	4	5	6	7	8	9
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49

$$
\begin{aligned}
&\min \xi^L,\\
\text{s.t. } &\left|w_B - a_{Bj} w_j\right| \le \xi^L, \forall j,\\
&\left|w_j - a_{jW} w_W\right| \le \xi^L, \forall j,\\
&\sum_j w_j = 1, w_j \ge 0, \forall j.
\end{aligned}
\tag{9}
$$

In this study, the classic linear BWM version of this method (Rezaei, 2016) was used, because it provides a unique solution, making it suitable for the comparison purposes in this study.

Consistency ratio of the results and its threshold to check the reliability of provided pairwise comparisons came from Liang et al. (2020). Equations 10 and 11 show the input-based consistency ratio CR^I for BWM comparisons.

$$
CR^I = \max_j CR_j^I, \tag{10}
$$

where,

$$
CR_j^I = \begin{cases} \dfrac{\left|a_{Bj} \times a_{jW} - a_{BW}\right|}{a_{BW} \times a_{BW} - a_{BW}}, & a_{BW} > 1,\\ 0, & a_{BW} = 1. \end{cases} \tag{11}
$$

CR^I is the global input-based consistency ratio for all criteria, while CR_j^I represents the local consistency level associated with the criterion c_j. Liang et al. (2020) obtained the consistency thresholds for combinations that range from 3 to 9 criteria (Table 3). For example, the thresholds in the combinations with 3-criteria and with 3-scale (scale is the largest evaluation grade from 3 to 9 in each comparison vector) is 0.1667. In this example, the provided pairwise comparison that its CR^I is lower than 0.1667 has an acceptable consistency.

4 | HYPOTHESES DEVELOPMENT

As stated in Section 2, equalizing bias refers to a situation where a decision-maker tends to assign (about) the same weight to all the decision-making attributes. Some researchers have argued that some weighting methods may be more prone to equalizing bias than others (Fox & Clemen, 2005; Marttunen et al., 2018; Tervonen et al., 2017), more specifically, that 'ranking events or objectives first, and then assigning ratio weights', is one of the debiasing strategies that can be used to mitigate this bias (Montibeller & von Winterfeldt, 2015a, 2015b, 2018). Methods like AHP and BWM are based on these mechanisms. On the other hand, researchers have also argued that the equal weight distribution of direct rating methods such as Swing, PA, and SMART is greater than in other methods (Marttunen et al., 2018; Pöyhönen & Hämäläinen, 2001; Tervonen et al., 2017). Moreover, we know that the procedures in some methods are based on the explicit pairwise comparison that compares relative importance of one attribute to that of other attributes in one or more turns by ratio scale (Rezaei, 2015, 2016; Saaty, 1977), assigning a ratio to each of the comparisons. Some of these methods rank attributes explicitly (for example identifying the best and worst attributes in BWM), while others (like AHP) do so implicitly. This makes it possible to compare the attributes, which ultimately leads to a greater distinction in the importance of attributes. However, methods like PA, SMART, and Swing use direct rating methods, making them more prone to equalizing biases (Tervonen et al., 2017). In other words, methods in which the importance is assigned directly to attributes (SMART, Swing, and PA) would suffer more from equalizing bias than those that use ratio scales (AHP and BWM). Based on these arguments, we want to test the following hypothesis:

> **Hypothesis 1.** AHP and BWM have less equalizing bias than SMART, Swing, and PA.

In addition, several researchers have proposed eliciting weights hierarchically as a debiasing strategy for mitigating the equalizing bias (Jacobi & Hobbs, 2007; Montibeller & von Winterfeldt, 2015a, 2015b, 2018; Sayeki & Vesper, 1973; Stillwell et al., 1987). Hierarchical weighting (Figure 1a) means that a decision-maker evaluates the attribute at each level and each cluster (the green circles compose one cluster at the first level; red circles and blue circles compose two clusters at the second level) of decision tree separately. The sum of the weights of each cluster is one. Global weights of the lowest level of the decision tree (blue and red attributes in Figure 1a) are then calculated by multiplying the weight of subattributes by their associated higher level attribute weight. The sum of weights of all global weights then becomes one. On the other hand, non-hierarchical weighting (Figure 1b) means that the decision-maker evaluates all the subattributes only and no evaluation is conducted for the higher levels

TABLE 3 Thresholds for consistency of a BWM problem with different combinations of criteria and scales (Liang et al., 2020)

Criteria							
Scales	**3**	**4**	**5**	**6**	**7**	**8**	**9**
3	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
4	0.1121	0.1529	0.1898	0.2206	0.2527	0.2577	0.2683
5	0.1354	0.1994	0.2306	0.2546	0.2716	0.2844	0.2960
6	0.1330	0.1990	0.2643	0.3044	0.3144	0.3221	0.3262
7	0.1294	0.2457	0.2819	0.3029	0.3144	0.3251	0.3403
8	0.1309	0.2521	0.2958	0.3154	0.3408	0.3620	0.3657
9	0.1359	0.2681	0.3062	0.3337	0.3517	0.3620	0.3662

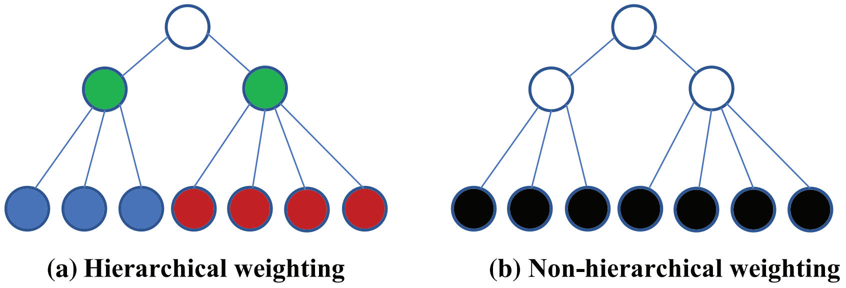
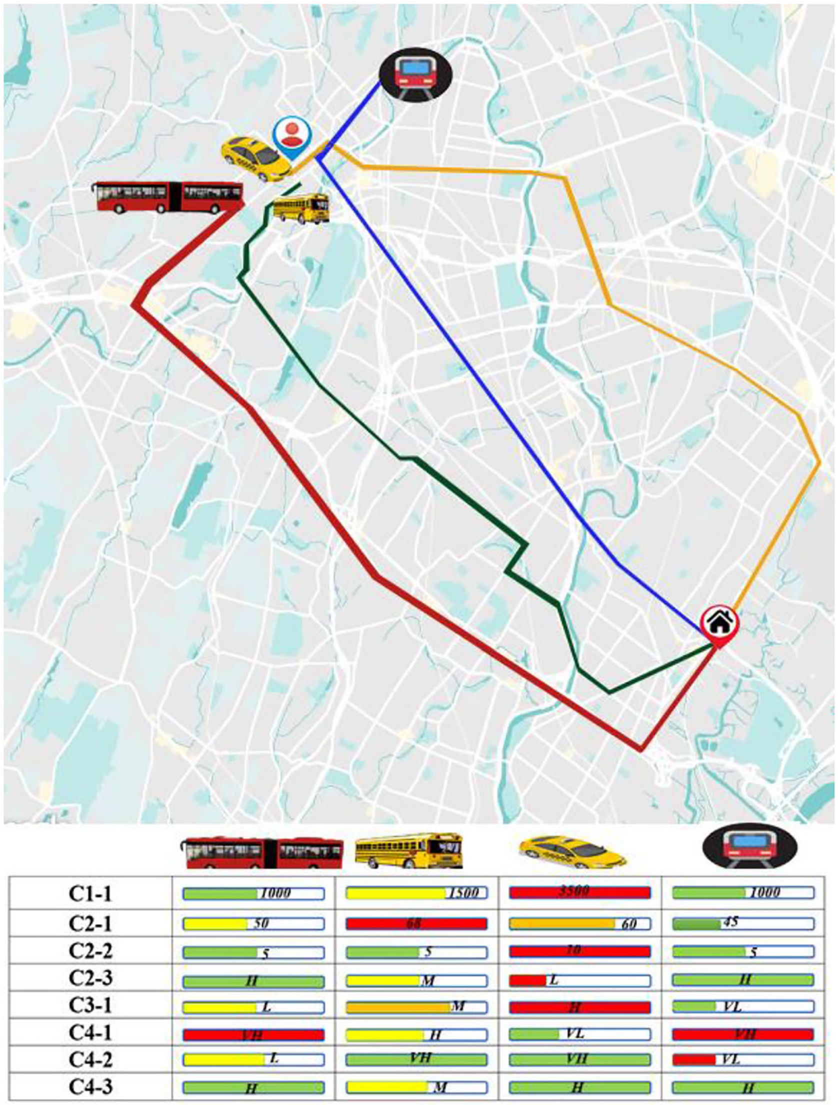


FIGURE 1 Hierarchical vs. non-hierarchical weighting [Colour figure can be viewed at wileyonlinelibrary.com]



C1-1	1000	1500	3500	1000
C2-1	50	68	60	45
C2-2	5	5	10	5
C2-3	H	M	L	H
C3-1	L	M	H	VL
C4-1	VH	H	VL	VH
C4-2	L	VH	VH	VL
C4-3	H	M	H	H

FIGURE 2 Location of lines and stations and the problem decision matrix [Colour figure can be viewed at wileyonlinelibrary.com]

of the hirarchy. It is evident that no further calculation is needed and the sum of the weights of all the subattributes (black circles in Figure 1b) is one.

In hierarchical weighting, the 1/*n* rule does not affect the entire set of attributes, because the decision-maker focuses on one cluster at a time, which could increase the chance of assigning nonequal weights to the attributes. However, with non-hierarchical weighting, the decision-maker focuses on all the subattributes, which now belong to one unified set and could increase the chance of being affected by the 1/*n* rule and equalizing bias. Also, researchers have argued that the number of objectives influences the distribution of weights due to the normalization of weights so that they add up to one (Marttunen et al., 2018; Pöyhönen & Hämäläinen, 1998, 2001; Weber et al., 1988). We know that non-hierarchical weighting has more objectives than hierarchical weighting (as in hierarchical the procedure is applied each time to one cluster), leading to more equalizing bias occurrence. Because of these arguments, we want to test the following hypothesis:

Hypothesis 2. The hierarchical structuring of the problem leads to a reduction in the equalizing bias in all five methods (AHP, BWM, PA, SMART, and Swing).

5 | EXPERIMENTAL STUDIES

5.1 | Decision problem scenario

This study uses a public transportation mode selection problem, the main aim being to weight the attributes and subattributes of the evaluation and selection of intra-city public transportation modes in Tehran. The problem is described to the participants as follows:

> On a Tuesday, with a normal temperate at 10 a.m., the subject plans to visit a friend who has already made an appointment with him/her. The origin and destination of the subject and the lines and stations' locations are fixed and showed in Figure 2. The decision matrix (the performance of each transportation mode with respect to each subattribute) considered as the input of the model. Also, subjects only have the right to choose one of the four mentioned modes of transport (Bus Rapid Transit (BRT), Bus, Taxi, and Metro).

For all subjects, both the transportation modes and the numbers in the decision (performance) matrix are the same. They are referred to in the same way in Figure 2. Besides, the attributes and subattributes of the problem (Table 4) are described for all subjects. These attributes and subattributes are extracted by a comprehensive literature review (for more details see Appendix A).

This problem was chosen to control some other biases as much as possible: (i) as a subject is imposed to a fixed and similar performance matrix for different tasks, range insensitivity (the sensitivity of the weights of the attributes to the changes of alternatives range) can be controlled, (ii) the problem has a fixed number of attributes and subattributes, and the structure is the same for all the tasks, which could control the splitting bias (presenting an attribute in greater detail [subattributes] and may increase the assigned weight), (iii) the attributes and subattributes of the problem were extracted by a comprehensive review of transportation literature, with the aim of making sure that all selected attributes are fundamental attributes that characterize different modes of transportation, and as such, the proxy bias (proxy attributes gain more weight than the fundamental attributes) is controlled, (iv) as we will discuss in more detail, the experimental design we choose for our study is a within-subjects design with randomization, which means that each subject can be treated as their own control.

Although it is not the focus of our study, we also tested the value function conditions for the attributes following Keeney and Raiffa (1976), Currim and Sarin (1984) and Dyer and Sarin (1979) using a sample of eight subjects, and it appeared that the two conditions of an additive value function hold for this example (mutual preferential independence and difference independence). We think that one of the main reasons behind having the two conditions is that the range of attributes in our case experiment is small (for more explanation, see von Winterfeldt and Edwards (1986) and Watson et al. (1987)).

TABLE 4 Attributes and subattributes of the research problem

Attribute	Subattribute	Subattribute description
Cost (C1)	Travel cost (C1–1)	Total payment for travel from origin to the final destination
Time (C2)	Travel time (C2–1)	The total time elapsed from the time the vehicle began to move until it reached its destination
	Waiting time (C2–2)	The total waiting time of the person at the station before the arrival and movement of the vehicle
	Reliability and punctuality of vehicles mode runs come on schedule to the destination (C2–3)	Nontime deviation of reaching the destination according to the pre-determined or expected plan for that vehicle
Environment friendly (C3)	Pollution (C3–1)	The amount of air pollution emitted by the vehicle
Comfort (C4)	The passenger density in the vehicle (C4–1)	Population and congestion within the vehicle
	Ease of accessibility to vehicle stop station (C4–2)	The ease and short distance to achieve the desired means of transportation
	Air condition and other equipment in the vehicles (C4–3)	Existence, use, and effectiveness of heating and cooling facilities in the vehicle

5.2 | Participants

Subjects selected from MSc students, MSc graduates and PhD candidates in the fields of Industrial Engineering, Management, and other fields in Tehran city familiar with MADM methods and willing to participate in the research. We selected respondents familiar with MADM methods to control the learning effect, which is one of the main concerns of within-subjects studies (Keren, 2014). In other words, the learning effect that would play a role with people who were not familiar with MADM could affect the reliability of the

TABLE 5 Subjects' characteristics ($n = 146$)

Characteristics	Levels/categories	Number (percent)
Education level	Master student	20 (13.7%)
	Masters	65 (44.5%)
	Ph.D. student	61 (41.8%)
Major	Management and industrial engineering	141 (96.6%)
	Miscellaneous (computer engineering, accounting, etc.)	5 (3.4%)
Age	[23,27)	21 (14.4%)
	[27,31)	61 (41.8%)
	≥31	64 (43.8%)
Gender	Male	81 (55.5%)
	Female	65 (44.5%)

findings. The channels for achieving these subjects were: public call in the LinkedIn, ResearchGate and then screening based on resume, taking advantage of the top national researchers in this field, entrusting them to the same subjects to introduce other subjects (snowball), and taking advantage of the opinions of professors at significant universities and introducing the subjects by them and the use of students is very common for this type of research (see, for instance, Buchanan and Corner (1997), Hämäläinen and Alaja (2008), Rezaei (2021)). Subjects participated voluntarily with no bonus for their participation, which reduced the chance of participation by unmotivated subjects (Rezaei, 2021). Although the necessary number of subjects for a reliable within-subjects experiment design is relatively small, due to its high-level controllability, to generate external validity, 158 questionnaires were sent to subjects, out of which 149 participants started the survey and 3 were dropped out due to incompleteness resulting in a collection of 146 correct and complete ones. This sample size is larger than similar studies such as Lin (2013), Hämäläinen and Alaja (2008), Pöyhönen and Hämäläinen (2000), Buchanan and Corner (1997), Fischer (1995), von Nitzsch and Weber (1993), and also larger than the sample size calculated by GPOWER 3.1[1] (2020) software for each statistical test we conduct in this study. Table 5 shows the characteristics of the subjects of this study.

5.3 | Experimental design

Generally speaking, there are two approaches in designing experiments: between-subjects design, and within-subjects design. While in the former case, each subject is assigned to a single treatment/task, in within-subjects design, each subject is assigned to all treatments/tasks. In this study we chose within-subjects design as it is more appropriate for the aim of our study. One of the main advantages of within-subjects design is that 'differences observed among conditions are not confounded with individual differences' (Keren, 2014, p. 258). Furthermore, it provides a greater degree of freedom, while at the same time making it possible to use a considerably smaller number of subjects in the experiment (Keren, 2014, p. 260). Within-subjects design can also be conducted in two different forms: either a subject is assigned to the same treatment/task several times, or a subject is never assigned to the same treatment/task more than once. In our study, we chose the latter option, because in that case, the subjects serve as their own control. Within-subjects design also has its drawbacks, however, one of the most important ones being the possible dependencies of the treatments/task, which could affect the findings. A possible remedy to handle such dependencies would be to randomize the order of treatments/tasks. In our study, we use counterbalancing (meaning that we created almost equal number of different order combinations) as our randomization method, and we used familiar subjects with MADM methods to reduce the learning effect. We also provided small examples in the beginning of each task which creates some time between conducting different tasks that could reduce the carryover effect (Greenwald, 1976).

5.4 | Response tasks and procedure

Subjects in this experiment completed all of the five MADM weighting methods (BWM, AHP, PA, SMART, and Swing) in hierarchical as well as non-hierarchical structures in a random sequence to minimize any possible carryover effect. We used the Gorilla platform (https://gorilla.sc) for data collection. This is one of the newest platforms for conducting experimental research virtually, which already has attracted many researchers (see, for instance, Daniel-Watanabe et al., 2020; Lavan et al., 2019; Love & Robinson, 2020). The main reasons for this choice are the existence of a flexible and comprehensive experiment design mechanism, questionnaire design, randomization mechanisms, data completion time storage, comprehensive management of subjects, and appropriate and easy user interface (Anwyl-Irvine et al., 2020). For a comprehensive study on the advantages of Gorilla in comparison with other platforms, see Anwyl-Irvine et al. (2020). Due to the specific application of MADM methods in the experimental research literature, it was impossible to construct specific questionnaires for each method in a pre-prepared manner. Therefore, the HTML programming language was used to overcome this limitation.

Subjects were randomly assigned (without replacement) in Gorilla to five methods questionnaires in hierarchical and non-hierarchical formats (10 tasks: 5 methods * 2 formats). We used counterbalancing method for randomization, which implies that we have an almost equal number of all possible order combinations of the methods and the two formats. The descriptions of each experiment task are provided briefly with a numerical example on each method page. There was no time limit for the subjects. In addition, the possibility of going back to the previous step in experimenting is disabled in all stages.

Each subject on average spent 43 min (s.d. = 25) finishing the entire experiment. We checked the average time subjects spent on different methods (in two hierarchical and non-hierarchical formats), which is as follows (all numbers are rounded). AHP: mean = 13 min (s.d. = 10); BWM: mean = 12 min (s.d. = 9); Swing: mean = 9 (s.d. = 8); SMART: mean = 5 (s.d. = 6); PA: mean = 4 (s.d. = 4). We think the difference between time spent on different methods, which are all statistically significantly different from each other ($p < 0.05$) is due to the amount of effort these methods require.

To check the subjects' level of tiredness, which can be partly reflected in the time they spend on a method when approaching the end of the experiment, we conducted some statistical tests to check the difference between the time subjects spend on a particular method when the method is their first randomly assigned test to subjects who have done that particular method as their fifth (last) randomly assigned method. We conducted the test for all five methods and found that the differences were not statistically significant, except for AHP. More specifically, those who performed AHP as their last method finished it relatively more quickly than those who performed it as their first method. This might be because the method is more popular, compared to other methods, amongst the respondents, which means that the time is mainly spent to understand the problem not the mechanism of the method. Therefore, those who were doing AHP as their last method had already become familiar with the problem and they were able to conduct the pairwise comparisons more quickly. This does not apply to the other methods. We proceeded to check the consistency ratio of those who performed AHP as their last method, to see if their consistency (on average) is different and we found that there is no statistically significant difference between the consistency ratio of the two groups, which means that them doing AHP more quickly at the end has no effect on the reliability of the weights.

TABLE 6 Pairwise comparisons of methods' equalizing bias for level 1

(I) Methods	(J) Methods	Equalizing bias index mean difference (I–J)	Std. error	Sig.[a]	95% confidence interval for difference[a] (lower bound)	95% confidence interval for difference[a] (upper bound)
AHP	BWM	0.005	0.005	1.000	−0.011	0.020
	PA	0.086**	0.005	0.000	0.071	0.102
	SMART	0.055**	0.005	0.000	0.041	0.069
	Swing	0.107**	0.005	0.000	0.092	0.122
BWM	PA	0.082**	0.006	0.000	0.066	0.097
	SMART	0.051**	0.005	0.000	0.037	0.064
	Swing	0.102**	0.005	0.000	0.087	0.118
PA	SMART	−0.031**	0.004	0.000	−0.043	−0.018
	Swing	0.021**	0.004	0.000	0.009	0.032
SMART	Swing	0.052**	0.004	0.000	0.040	0.063

[a] Adjustment for multiple comparisons: Bonferroni.

$^{*}p < 0.05$;

$^{**}p < 0.005$.

TABLE 7 Pairwise comparisons of methods' equalizing bias for level 2

(I) Methods	(J) Methods	Equalizing bias index mean difference (I–J)	Std. error	Sig.[a]	95% confidence interval for difference[a] (lower bound)	95% confidence interval for difference[a] (upper bound)
AHP	BWM	0.010**	0.003	0.003	0.002	0.018
	PA	0.048**	0.002	0.000	0.041	0.055
	SMART	0.034**	0.002	0.000	0.027	0.041
	Swing	0.053**	0.002	0.000	0.047	0.060
BWM	PA	0.038**	0.002	0.000	0.031	0.045
	SMART	0.024**	0.002	0.000	0.017	0.030
	Swing	0.043**	0.002	0.000	0.036	0.050
PA	SMART	−0.015**	0.002	0.000	−0.020	−0.009
	Swing	0.005*	0.002	0.015	0.001	0.009
SMART	Swing	0.019**	0.002	0.000	0.014	0.025

[a] Adjustment for multiple comparisons: Bonferroni.

$^{*}p < 0.05$;

$^{**}p < 0.005$.

6 | DATA ANALYSIS AND DISCUSSION

This section contains the analysis based on the collected data and the calculated normalized weights of each method. In this way, after collecting completed experiment tasks, all of the five methods in two problem structure formats solved by Excel solver that was developed especially for this experiment and global weight of all subattributes of each method calculated for each subject. Then a bias index is required to analyze the equalizing bias occurrence in these methods, which is defined in our study as follows.

Definition 4. The equalizing bias of participant s following method m, e_s^m, is calculated using the following formula:

$$e_s^m = \sqrt{\frac{\sum_{j=1}^{n}\left(w_{js}^m - \frac{1}{n}\right)^2}{n}}, \forall s \in S, m \in M, \tag{12}$$

where, w_{js}^m is the weight of attribute j ($j = 1,2, \ldots, n$) for participant s ($s = 1, 2, \ldots, S$) in the method m ($m = 1, 2, \ldots, M$); $\frac{1}{n} = \overline{w}_s^m$ is the mean of attributes weight in the method m for participant s (as the number of attributes is equal for all methods and all participants and as the sum of weights is one, $\overline{w}_s^m = \frac{1}{n}$).

This bias index is similar to the standard deviation in statistics and plays a vital role in almost all statistical inference procedures, especially measures of variability. A low bias index (close to 0) indicates that the weights tend to be close to the mean (high equalizing bias incidence). By contrast, a high bias index indicates that the weights are spread out over a wider range (low equalizing bias incidence).

After calculating the equalizing bias index for each method, all of these data were analyzed with SPSS version 26.0 to test the hypotheses listed in Section 4.

6.1 | Test of Hypothesis 1

To test Hypothesis 1, repeated measures analysis of variance (ANOVA) was used to examine differences for each bias index e_s^m for the two levels of the decision tree. During the initial testing of assumptions, Mauchly's test of sphericity indicated that the assumption of sphericity had not been met for the methods' effect on the equalizing bias index ($\chi^2 \geq 36.47$, $p < 0.05$) for level 1 and ($\chi^2 \geq 64.12$, $p < 0.05$) for level 2. Therefore, the Greenhouse–Geisser correction was used to calculate a conservative comparison of equalizing bias index means for both levels. Bonferroni post hoc analyses were conducted to determine the ranking of methods' bias. A test of within-subjects effects shows that there was a significant main effect of methods on equalizing bias index ($F(3.56,\ 516.74) = 185.77$, $p < 0.05$) for level 1 and ($F(3.381,\ 490.28) = 213.60$, $p < 0.05$) for level 2.

Hence, to identify the exact differences of the above findings, Bonferroni post hoc analyses were conducted. The results show that there are statistically significant differences between the equalizing bias index means at the first level except for AHP and BWM (Table 6) and for equalizing bias index of all methods for level 2 (Table 7). Also, ranking the methods by equalizing bias index means for the two levels (Table 8, Table 9 and Figure 3) shows that AHP (equalizing bias mean: 0.207 for level 1 and 0.116 for level 2), BWM (equalizing bias mean: 0.202 for level 1 and 0.106 for level 2), SMART (equalizing bias mean: 0.151 for level 1 and 0.083 for level 2), PA (equalizing bias

TABLE 8 Estimates of methods' equalizing bias for level 1

Methods	Equalizing bias index mean	Std. error	95% confidence interval for difference (lower bound)	95% confidence interval for difference (upper bound)
AHP	0.207	0.005	0.198	0.216
BWM	0.202	0.005	0.193	0.212
PA	0.121	0.005	0.112	0.130
SMART	0.151	0.003	0.145	0.158
Swing	0.100	0.004	0.092	0.108

TABLE 9 Estimates of methods' equalizing bias for level 2

Methods	Equalizing bias index mean	Std. error	95% confidence interval for difference (lower bound)	95% confidence interval for difference (upper bound)
AHP	0.116	0.002	0.112	0.121
BWM	0.106	0.002	0.102	0.111
PA	0.068	0.002	0.064	0.072
SMART	0.083	0.002	0.079	0.086
Swing	0.063	0.001	0.060	0.066

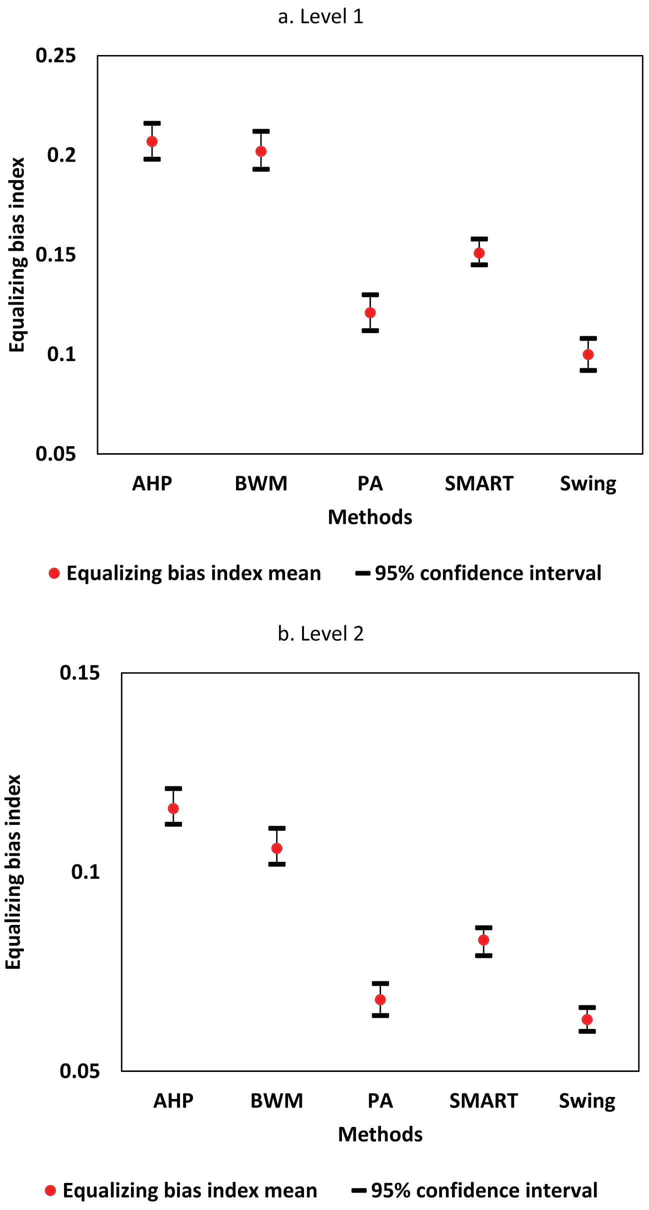


FIGURE 3 Estimated marginal means of methods' equalizing bias [Colour figure can be viewed at wileyonlinelibrary.com]

mean: 0.121 for level 1 and 0.068 for level 2), and Swing (equalizing bias mean: 0.1 for level 1 and 0.063 for level 2) have less to more equalizing bias, respectively, which means that this hypothesis is supported.

AHP and BWM displayed less equalizing bias than SMART, Swing, and PA, which was in line with the debiasing strategy presented by Montibeller and von Winterfeldt (2015b), Montibeller and von Winterfeldt (2015a), and Montibeller and von Winterfeldt (2018) to use 'rank and ratio-based methods' in weighting the attributes. These methods are working based on pairwise comparison that compares the relative importance of one attribute to other attributes in one or more turns and assigning a ratio to each of the comparisons. Despite several criticisms to AHP (see, for instance, Barzilai (1997) for 'rank reversal', Salo and Hämäläinen (1997) for 'judgement scales', and (Dyer, 1990) for 'accuracy of results'), not directly assigning importance scores to the attributes by this method and comparing attributes to each other results in a more significant distinction in the relative importance of the attributes. In BWM, the use of pairwise comparison also leads to more distinction in the weights.

In BWM, the decision-maker explicitly ranks the attributes in Best-to-Others and Others-to-Worst vectors. Although this is not done explicitly in AHP, the decision-maker needs to consider the ranking of the attributes implicitly, to make a pairwise comparison of each attribute with the other attributes. On the other hand, despite the initial ranking procedure of attributes in SMART and Swing methods, these methods, together with the PA method, are direct rating methods in which the decision-maker directly assigns the importance of each attribute and does not compare the relative importance of the attributes, making them more prone to equalizing, as indicated in some papers, including Marttunen et al. (2018), Pöyhönen and Hämäläinen (2001) and Tervonen et al. (2017).

6.2 | Test of Hypothesis 2

Here, we first tested the effect of problem structuring (hierarchical, non-hierarchical) on equalizing bias, considering all methods

TABLE 10 Pairwise comparisons of the structure of the problem's equalizing bias

(I) Structure of weighting	(J) Structure of weighting	Equalizing bias index mean difference (I–J)	Std. error	Sig.[a]	95% confidence interval for difference[a] (lower bound)	95% confidence interval for difference[a] (upper bound)
Hierarchical	Non-hierarchical	0.027**	0.002	0.000	0.024	0.031

[a] Adjustment for multiple comparisons: Bonferroni.
*p < 0.05;
**p < 0.005.

TABLE 11 Estimates of the structure of the problem's equalizing bias

Structure of weighting	Equalizing bias index mean	Std. error	95% confidence interval for difference (lower bound)	95% confidence interval for difference (upper bound)
Hierarchical	0.101	0.002	0.097	0.105
Non-hierarchical	0.074	0.001	0.071	0.076

together, after which we examined the difference between the various methods in this respect. Initially, repeated measures ANOVA was used to examine differences for each bias index. Bonferroni post hoc analyses were conducted to determine the ranking of the problem structuring formats bias. A test of within-subjects effects (considering all five methods together) shows that there was a significant main effect of hierarchical/non-hierarchical structure of problem on equalizing bias ($F(1.145) = 234.59$, $p < 0.05$).

Hence, to determine the exact differences in the findings outlined, Bonferroni post hoc analyses were conducted. The results show that there was a significant difference between the hierarchical/non-hierarchical structuring of the problem in equalizing bias index (mean difference 0.027, $p < 0.05$) (Table 10), proving that hierarchical structuring (equalizing bias mean: 0.101) is statistically significantly less prone to equalizing bias than non-hierarchical structure (equalizing bias mean: 0.074) (Table 11 and Figure 4).

The hierarchical structuring of the problem leads to a reduction in the occurrence of equalizing bias, which was in line with the debiasing strategy presented by Montibeller and von Winterfeldt (2018), Montibeller and von Winterfeldt (2015a), Montibeller and von Winterfeldt (2015b), Stillwell et al. (1987), where the use of a hierarchical approach in weighting the attributes leads to fewer occurrences of equalizing bias. Hierarchical weighting is a comparison of attributes/subattributes that are on a similar level of a decision tree and in the same cluster. This leads to comparing attributes in smaller subsets and finally mitigate the splitting bias (Hämäläinen & Alaja, 2008), because the distribution of weights is

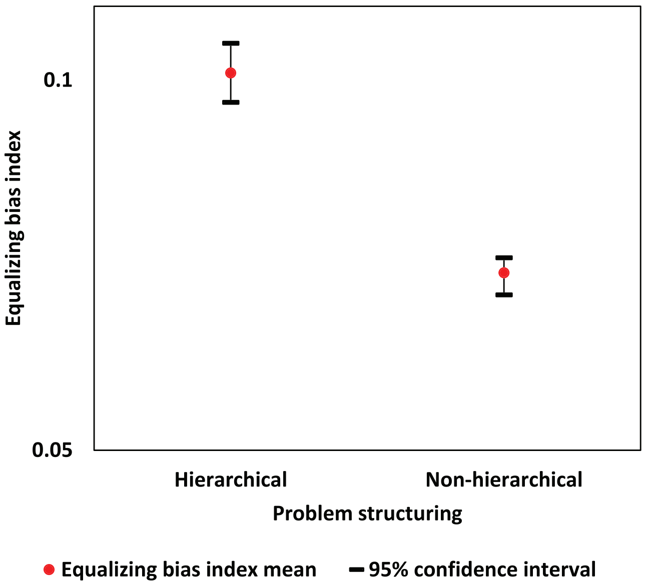


FIGURE 4 Estimated marginal means of the structure of the problem's equalizing bias [Colour figure can be viewed at wileyonlinelibrary.com]

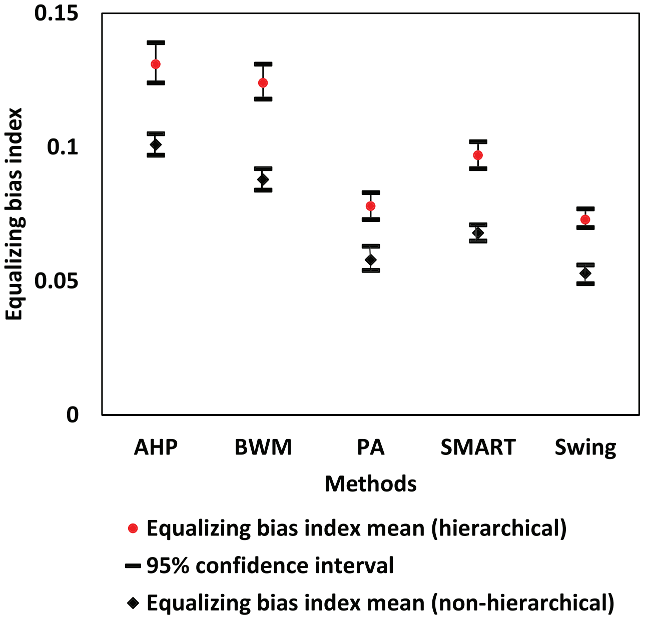


FIGURE 5 Estimated marginal means of the interaction effect of the structuring of the problem on the equalizing bias among all five methods [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 12 The interaction effect of the structuring of the problem on the equalizing bias among all five methods estimates

Methods	Structure of the problem	Equalizing bias index mean	Std. error	95% confidence interval for difference (lower bound)	95% confidence interval for difference (upper bound)
AHP	Hierarchical	0.131	0.004	0.124	0.139
	Non-hierarchical	0.101	0.002	0.097	0.105
BWM	Hierarchical	0.124	0.003	0.118	0.131
	Non-hierarchical	0.088	0.002	0.084	0.092
PA	Hierarchical	0.078	0.002	0.073	0.083
	Non-hierarchical	0.058	0.002	0.054	0.063
SMART	Hierarchical	0.097	0.002	0.092	0.102
	Non-hierarchical	0.068	0.001	0.065	0.071
Swing	Hierarchical	0.073	0.002	0.070	0.077
	Non-hierarchical	0.053	0.002	0.049	0.056

influenced by the number of objectives due to the normalization of weights, so that they add up to one (Marttunen et al., 2018; Pöyhönen & Hämäläinen, 1998, 2001; Weber et al., 1988). Subjects have a response bias against large weight ratios. Since relatively flat higher level weights can still produce steep lower level weights, this response bias would affect non-hierarchical weights more than hierarchical weights and leads to flatter weights in the non-hierarchical structure (Stillwell et al., 1987).

Furthermore, to test the interaction effect of the hierarchical structuring of the problem on equalizing bias among all five methods, repeated measures ANOVA was used to examine differences for each bias measure. During the initial testing of assumptions, Mauchly's test of sphericity indicated that the assumption of sphericity had not been met for the effect of the structuring of the problem on the equalizing bias index among all five methods ($\chi^2 \geq 64.12$, $p < 0.05$). Therefore, the Greenhouse–Geisser correction was used to calculate a conservative comparison of the equalizing bias index means. Bonferroni post hoc analyses were conducted to determine the ranking of methods formats' bias. A test of within-subjects effects shows that there was a significant main effect of the structuring of the problem on the equalizing bias among all five methods ($F(3.24, 470.17) = 10.19$, $p < 0.05$).

Hence, to determine the exact differences in the findings outlined above, Bonferroni post hoc analyses were conducted. The ranking of methods in each problem structuring by equalizing bias index means show (Table 12 and Figure 5) the hierarchical AHP (equalizing bias index mean: 0.131), hierarchical BWM (equalizing bias index mean: 0.124), non-hierarchical AHP (equalizing bias index mean: 0.101), hierarchical SMART (equalizing bias index mean: 0.097), non-hierarchical BWM (equalizing bias index mean: 0.088), hierarchical PA (equalizing bias index mean: 0.078), hierarchical Swing (equalizing bias index mean: 0.073), non-hierarchical SMART (equalizing bias index mean: 0.068), non-hierarchical PA (equalizing bias index mean: 0.058), and non-hierarchical Swing (equalizing bias index mean: 0.053) demonstrated from less to more equalizing bias, respectively.

In addition, the paired samples *t* test was used to examine differences for the effect of hierarchical viewing of the problem on the equalizing bias in each method. Similar to the results of repeated measures ANOVA, the results show that all methods equalizing bias index significantly differ in view of problem structure. The hierarchical problem structuring leads to a reduction in equalizing bias in all the five methods, although to a significantly different degree. The BWM and AHP show more hierarchical differences from the non-hierarchical structure (0.0363 and 0.0300, respectively) than the other methods (Table 13). It can be said that, the lower the occurrence of equalizing bias in a method, the greater the difference of this bias index in the hierarchical and non-hierarchical structure of the same method.

6.3 | Implications of our findings for splitting bias

This section discusses an additional finding of this study. A part of Section 6.1 focused on checking the equalizing bias for the first level of the tree for the five different methods. We conducted the analysis based on the weights obtained by hierarchical weighting. It is also possible to calculate the weights of the first level attributes by adding the weights of the subattributes of a main attribute when the weights have been calculated non-hierarchically. In this section, we calculate the weights this way and check the equalizing bias for the methods. Following the same approach as in Section 6.1, we found the equalizing bias index means for the five methods that are presented in Table 14. As can be seen, this approach yields different results, in particular for PA, SMART, and Swing. We think that this is related to another bias called splitting bias, which has been studied by Borcherding and von Winterfeldt (1988), Jacobi and Hobbs (2007), Pöyhönen and Hämäläinen (1998), Weber et al. (1988) and others. It is defined as 'the phenomenon where dividing an existing attribute into subattributes in a branch of a value tree produces an increase in the overall weight of that branch when non-hierarchical weighting is used' (Hämäläinen & Alaja, 2008). Experimental studies on splitting

TABLE 13 Effect of hierarchical viewing of the problem on the equalizing bias in each method

Pair	Equalizing bias index mean	Std. deviation	Std. error mean (paired differences)	95% confidence interval for paired difference (lower bound)	95% confidence interval for paired difference (upper bound)	*t*	Df	Sig. (2-tailed)
AHP hierarchical—AHP non-hierarchical	0.030**	0.040	0.003	0.024	0.036	9.166	145	0.000
BWM hierarchical—BWM non-hierarchical	0.036**	0.035	0.003	0.031	0.042	12.552	145	0.000
PA hierarchical—PA non-hierarchical	0.020**	0.029	0.002	0.015	0.024	8.243	145	0.000
SMART hierarchical—SMART non-hierarchical	0.030**	0.029	0.002	0.025	0.034	12.471	145	0.000
Swing hierarchical—Swing non-hierarchical	0.021**	0.027	0.002	0.016	0.025	9.206	145	0.000

*$p < 0.05$; **$p < 0.005$.

TABLE 14 Effect of hierarchical and non-hierarchical viewing of the problem on the equalizing bias of first level attributes in each method

Pair	Equalizing bias index mean	Std. deviation	Std. error mean (paired differences)	95% confidence interval for paired difference (lower bound)	95% confidence interval for paired difference (upper bound)	*t*	Df	Sig. (2-tailed)
AHP hierarchical—AHP non-hierarchical	−0.021**	0.075	0.00617	−0.033	−0.008	−3.351	145	0.001
BWM hierarchical—BWM non-hierarchical	−0.005	0.058	0.00478	−0.014	0.005	−0.987	145	0.325
PA hierarchical—PA non-hierarchical	−0.075**	0.043	0.00359	−0.082	−0.068	−20.845	145	0.000
SMART hierarchical—SMART non-hierarchical	−0.359**	0.107	0.00889	−0.377	−0.342	−40.420	145	0.000
Swing hierarchical—Swing non-hierarchical	−0.419**	0.088	0.00725	−0.433	−0.404	−57.766	145	0.000

$^{*}p < 0.05$;
$^{**}p < 0.005$.

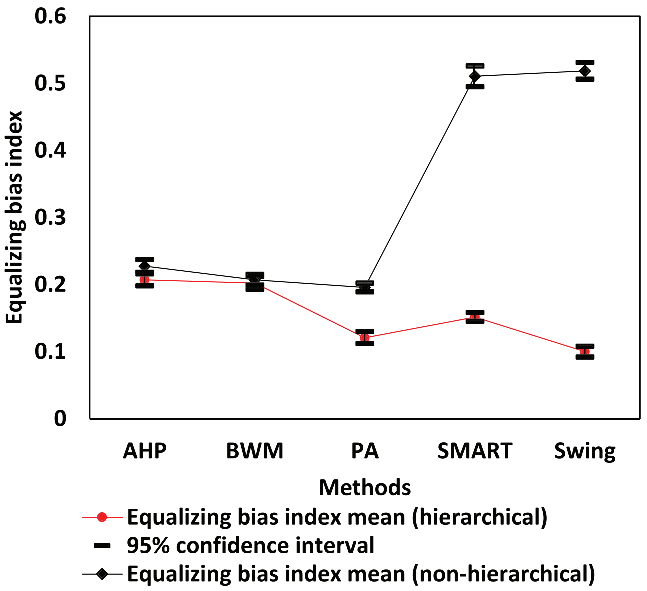


FIGURE 6 Equalizing bias among all five methods for the five different methods for the first level attributes' weight (hierarchical vs. non-hierarchical problem structuring) [Colour figure can be viewed at wileyonlinelibrary.com]

bias are limited to Swing (Borcherding & von Winterfeldt, 1988; Hämäläinen & Alaja, 2008; Pöyhönen et al., 2001; Pöyhönen & Hämäläinen, 2000; Weber et al., 1988), PA (Jacobi & Hobbs, 2007), Tradeoff (Borcherding & von Winterfeldt, 1988), Pricing out (Borcherding & von Winterfeldt, 1988), Otto (Weber et al., 1988), Conjoint (Weber et al., 1988), and Ratio (Borcherding & von Winterfeldt, 1988; Weber et al., 1988). If we compare the index bias for the first-level attributes using the two hierarchical and non-hierarchical weighting, we see very interesting results (see Table 14 and Figure 6). As can be seen, the difference between equalizing bias indexes of the two hierarchical and non-hierarchical weightings for PA, SMART, and Swing are statistically significantly different. For AHP the difference, although significant, the equalizing bias indexes are not too far from each other and for BWM they are not statistically significantly different and as we see in Figure 6, they have a significant overlap. Based on this observation, we would conclude that methods that are more prone to equalizing bias are also more prone to splitting bias, and that splitting bias is not necessarily always a phenomenon in the case of non-hierarchical weighting, for instance, when using BWM. However, our findings confirm that splitting bias occurs with non-hierarchical weighting, but not for all methods, so we think more research is needed in this area.

7 | CONCLUSION AND FUTURE RESEARCH

Attribute weighting is one of the most important steps in a multiattribute decision-making (MADM) problem. Most MADM weighting methods are based on the assessments of experts/decision-makers, which are prone to several cognitive biases and lead to suboptimal results. Therefore, it is necessary to examine these biases to improve the methods and develop debiasing strategies. To date, however, few studies have addressed this issue from an analytical-experimental perspective. Our study is one of a few experimental studies to examine the existence of the equalizing bias in MADM weighting methods.

In this research, five MADM weighting methods AHP, BWM, SMART, Swing, and PA, under two structuring formats, hierarchical and non-hierarchical, were selected for the experiment. The hypotheses were developed and then tested by an experiment design. We found that the hierarchical structuring of the problem leads to a significant reduction of equalizing bias in all methods under examination, a result that is in line with Jacobi and Hobbs (2007), Montibeller and von Winterfeldt (2015b), Montibeller and von Winterfeldt (2015a), Montibeller and von Winterfeldt (2018), Sayeki and Vesper (1973) and Stillwell et al. (1987), who suggested hierarchical weighting as a strategy for eliciting the weight of the attributes equally. Also, this result is indirectly in line with Pöyhönen and Hämäläinen (2001), Pöyhönen and Hämäläinen (1998), Marttunen et al. (2018) and Weber et al. (1988) who described the 1/*n* rule and the number of attributes as a factor affecting the distribution of the weights. In this way, the hierarchical structuring of the problem has fewer attributes to be considered together, because the attributes are clustered in smaller subsets (compared to non-hierarchical situation, where all sub-attributes are evaluated together) and reduces equalizing bias. In addition, AHP and BWM have less equalizing bias than SMART, Swing, and PA, which proves the efficiency of the debiasing strategies that have been proposed in the existing studies to use 'rank and ratio-based methods' in weighting the attributes. This result was in line with Montibeller and von Winterfeldt (2015b), Montibeller and von Winterfeldt (2015a), and Montibeller and von Winterfeldt (2018). It was also in line with Pöyhönen and Hämäläinen (2001), Tervonen et al. (2017), and Marttunen et al. (2018), who suggested direct rating methods such as Swing, PA, and SMART lead to equal weight distribution more than other methods. The findings suggest that SMART, Swing, and PA are more suitable for hierarchical structures (than non-hierarchical structures) and that if the questions being asked in these methods are changed to some explicit ratio questions, the conclusions may be less prone to equalizing bias. The findings also help researchers consider debiasing strategies to counter equalizing bias when developing new methods.

This study has a number of limitations. First of all, we collected the data via a virtual platform, but feel that collecting data in a laboratory setting would be more desirable, because in that case, we could better control the condition of the experiments. The experiment we designed was a within-subjects design with ten different tasks, which is a bit long, and subjects may have become tired towards the end of the experiment, based on which we recommend reducing the number of tasks for the participants. In an extreme scenario, the experiment can be designed based on between-subject design, which has its own advantages and disadvantages (e.g., a main disadvantage is that it requires more participants, while a main advantage is that it is naturally not affected by learning effect, carryover effect, and dependency between the treatments/tasks). As argued above, choosing between within-subjects design and between-subjects design is in itself a multiattribute decision-making problem! Finally, while, like almost all earlier studies in this area, we focused on one cognitive bias (equalizing bias), and we think that some of biases may be interconnected and that future research could examine such possible connections. For instance, we think that equalizing bias and anchoring bias may have some associations, for instance, when a weighting method could better handle the anchoring bias (for instance by using multiple anchors), it is most likely less prone to equalizing bias too. We think that, if we identify such a relationship among cognitive biases in MADM weighting methods, we could better formulate debiasing strategies, because when we suggest debiasing strategies for a particular cognitive bias and ignore all other biases, such a debiasing strategy may not be very effective. Finally, while we identified equalizing bias in MADM weighting methods in an experimental setting, as the real decision-making involves a close interaction between the analyst and the decision-maker, the analyst could help the decision-maker overcome such a bias, which can a subject for future research.

ENDNOTE

[1] One of the most common and powerful software programs for estimating the number of subjects required for research based on statistical tests (Faul et al., 2007, 2009)

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author, upon request.

ORCID

Jafar Rezaei https://orcid.org/0000-0002-7407-9255
Alireza Arab https://orcid.org/0000-0002-5024-2177
Mohammadreza Mehregan https://orcid.org/0000-0002-7974-8171

REFERENCES

Alaja, S., & Hämäläinen, R. P. (2008). The threat of weighting biases in environmental decision analysis. *Ecological Economics*, *68*(1–2), 556–569. https://doi.org/10.1016/j.ecolecon.2008.05.025

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Barzilai, J. (1997). Deriving weights from pairwise comparison matrices. *Journal of the Operational Research Society*, *48*(12), 1226–1232. https://doi.org/10.1057/palgrave.jors.2600474

Bazerman, M. H., & Moore, D. A. (2012). *Judgment in managerial decision making*. 8, Wiley.

Belton, V., Lienert, J., & Marttunen, M. (2018). Are objectives hierarchy related biases observed in practice? A meta-analysis of environmental and energy applications of multi-criteria decision analysis. *European Journal of Operational Research*, *265*(1), 178–194. https://doi.org/10.1016/j.ejor.2017.02.038

Borcherding, K., & von Winterfeldt, D. (1988). The effect of varying value trees on multiattribute evaluations. *Acta Psychologica*, *68*(1–3), 153–170. https://doi.org/10.1016/0001-6918(88)90052-2

Borcherding, K., & Weber, M. (1993). Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research*, *67*(1), 1–12. https://doi.org/10.1016/0377-2217(93)90318-H

Bottomley, P. A., & Doyle, J. R. (2001). A comparison of three weight elicitation methods: Good, better, and best. *Omega*, *29*(6), 553–560. https://doi.org/10.1016/S0305-0483(01)00044-5

Bottomley, P. A., Doyle, J. R., & Green, R. H. (1997). Judging relative importance: Direct rating and point allocation are not equivalent. *Organizational Behavior and Human Decision Processes*, *70*(1), 65–72. https://doi.org/10.1006/obhd.1997.2694

Brunelli, M., Liang, F., & Rezaei, J. (2020). Consistency issues in the best worst method: Measurements and thresholds. *Omega*, *96*, 102175. https://doi.org/10.1016/j.omega.2019.102175

Buchanan, J. T., & Corner, J. (1997). The effects of anchoring in interactive MCDM solution methods. *Computers & Operations Research*, *24*(10), 907–918. https://doi.org/10.1016/S0305-0548(97)00014-2

Buede, D. M., & Watson, S. R. (1987). *Decision synthesis: The principles and practice of decision analysis*. Cambridge Univ. Press.

Clemen, R. T., & Fox, C. R. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, *51*(9), 1417–1432. https://doi.org/10.1287/mnsc.1050.0409

Currim, I. S., & Sarin, R. K. (1984). A comparative evaluation of multiattribute consumer preference models. *Management Science*, *30*(5), 543–561. https://doi.org/10.1287/mnsc.30.5.543

Damodaran, N., Fischer, G. W., Laskey, K. B., & Lincoln, D. (1987). Preferences for proxy attributes. *Management Science*, *33*(2), 198–214. https://doi.org/10.1287/mnsc.33.2.198

Daniel, L., Gormley, S., McLaughlin, M., & Robinson, O. J. (2020). Association between a directly translated cognitive measure of negative bias and self-reported psychiatric symptoms. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. In press. https://doi.org/10.1016/j.bpsc.2020.02.010

Dantzig, G. B., Hogarth, R., Plott, C. R., Raiffa, H., Schelling, T. C., Shepsle, K. A., Simon, H. A., Thaler, R., Tversky, A., & Winter, S. (1987). Decision making and problem solving. *INFORMS Journal on Applied Analytics*, *17*(5), 11–31. https://doi.org/10.1287/inte.17.5.11

Dyer, J. S. (1990). Remarks on the analytic hierarchy process. *Management Science*, *36*(3), 249–258. https://doi.org/10.1287/mnsc.36.3.249

Dyer, J. S., & Sarin, R. K. (1979). Measurable multiattribute value functions. *Operations Research*, *27*(4), 810–822. https://doi.org/10.1287/opre.27.4.810

Edwards, W. (1977). How to use multiattribute utility measurement for social decisionmaking. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, *7*(5), 326–340. https://doi.org/10.1109/TSMC.1977.4309720

Edwards, W. & von Winterfeldt, D. (1986). Decision analysis and behavioral research: Cambridge Univ. Press.

Eisenführ, F., Weber, M., & von Winterfeldt, D. (1988). The effects of splitting attributes on weights in multiattribute utility measurement. *Management Science*, *34*(4), 431–445. https://doi.org/10.1287/mnsc.34.4.431

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175–191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149–1160.

Fischer, G. W. (1995). Range sensitivity of attribute weights in multiattribute value models. *Organizational Behavior and Human Decision Processes*, *62*(3), 252–266. https://doi.org/10.1006/obhd.1995.1048

Gabrielli Jr, W. F., & von Winterfeldt, D. (1978). *Are Important Weights Sensitive to the Range of Alternatives in Multiattribute Utility Measurement*. Decisions and Designs Inc, Mclean,VA, 78(6), 1–30.

Gelhorn, H., Gries, K. S., Rentz, A., Marsh, K., Poon, J. L., Sri Bhashyam, S., & Tervonen, T. (2017). MCDA swing weighting and discrete choice experiments for elicitation of patient benefit-risk preferences: A critical assessment. *Pharmacoepidemiology and Drug Safety*, *26*(12), 1483–1491. https://doi.org/10.1002/pds.4255

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*(2), 314–320. https://doi.org/10.1037/0033-2909.83.2.314

Hämäläinen, R. P. (2015). Behavioural issues in environmental modelling—The missing perspective. *Environmental Modelling & Software*, *73*, 244–253. https://doi.org/10.1016/j.envsoft.2015.08.019

Hämäläinen, R. P., Jenytin, C., & Lahtinen, T. J. (2020). On preference elicitation processes which mitigate the accumulation of biases in multi-criteria decision analysis. *European Journal of Operational Research*, *282*(1), 201–210. https://doi.org/10.1016/j.ejor.2019.09.004

Hämäläinen, R. P., & Pöyhönen, M. (1998). Notes on the weighting biases in value trees. *Journal of Behavioral Decision Making*, *11*(2), 139–150. https://doi.org/10.1002/(SICI)1099-0771(199806)11:2<139::AID-BDM293>3.0.CO;2-M

Hämäläinen, R. P., & Pöyhönen, M. (2000). There is hope in attribute weighting. *INFOR: Information Systems and Operational Research*, *38*(3), 272–282. https://doi.org/10.1080/03155986.2000.11732412

Hämäläinen, R. P., & Pöyhönen, M. (2001). On the convergence of multiattribute weighting methods. *European Journal of Operational Research*, *129*(3), 569–585. https://doi.org/10.1016/S0377-2217(99)00467-1

Hämäläinen, R. P., Pöyhönen, M., & Vrolijk, H. (2001). Behavioral and procedural consequences of structural variation in value trees. *European Journal of Operational Research*, *134*(1), 216–227. https://doi.org/10.1016/S0377-2217(00)00255-1

Hämäläinen, R. P., & Salo, A. A. (1997). On the measurement of preferences in the analytic hierarchy process. *Journal of Multi-Criteria Decision Analysis*, *6*(6), 309–319. https://doi.org/10.1002/(SICI)1099-1360(199711)6:6<309::AID-MCDA163>3.0.CO;2-2

Hobbs, B. F., & Jacobi, S. K. (2007). Quantifying and mitigating the splitting bias and other value tree-induced weighting biases. *Decision Analysis*, *4*(4), 194–210. https://doi.org/10.1287/deca.1070.0100

John, R. S., Stillwell, W. G., & von Winterfeldt, D. (1987). Comparing hierarchical and nonhierarchical weighting methods for eliciting multiattribute value models. *Management Science*, *33*(4), 442–450. https://doi.org/10.1287/mnsc.33.4.442

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. 10.1017/CBO9780511809477

Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives*. Wiley.

Keren, G. (2014). Between-or within-subjects design: A methodological dilemma. *A Handbook for Data Analysis in the Behaviorial Sciences*, *1*, 257–272.

Knight, S., Lavan, N., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 1–9.

Korhonen, P., Pajala, T., & Wallenius, J. (2019). Judgments of importance revisited: What do they mean? *Journal of the Operational Research Society*, *70*(7), 1140–1148. https://doi.org/10.1080/01605682.2018.1489346

Lin, S.-W. (2013). An investigation of the range sensitivity of attribute weight in the analytic hierarchy process. *Journal of Modelling in Management*, *8*(1), 65–80. https://doi.org/10.1108/17465661311311987

Love, J., & Robinson, O. J. (2020). "Bigger" or "better": The roles of magnitude and valence in "affective bias". *Cognition and Emotion*, *34*(4), 633–642. https://doi.org/10.1080/02699931.2019.1662373

Montibeller, G. (2018). Behavioral challenges in policy analysis with conflicting objectives. In *Recent advances in optimization and modeling of contemporary problems* (pp. 85–108). INFORMS.

Montibeller, G., & von Winterfeldt, D. (2015a). Biases and debiasing in multi-criteria decision analysis. Paper presented at the 2015 48th Hawaii International Conference on System Sciences, 5-8 January 2015, Kauai, Hawaii, USA, pp. 1218–1226.

Montibeller, G., & von Winterfeldt, D. (2015b). Cognitive and motivational biases in decision and risk analysis. *Risk Analysis*, *35*(7), 1230–1251. https://doi.org/10.1111/risa.12360

Montibeller, G., & von Winterfeldt, D. (2018). Individual and group biases in value and uncertainty judgments. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation* (Vol. 261) (pp. 377–392). Springer.

Pennings, J. M., van Ittersum, K., van Trijp, H. C., & Wansink, B. (2007). The validity of attribute-importance measurement: A review. *Journal of Business Research*, *60*(11), 1177–1190. https://doi.org/10.1016/j.jbusres.2007.04.001

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57. https://doi.org/10.1016/j.omega.2014.11.009

Rezaei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega*, *64*, 126–130. https://doi.org/10.1016/j.omega.2015.12.001

Rezaei, J. (2021). Anchoring bias in eliciting attribute weights and values in multi-attribute decision-making. *Journal of Decision Systems*, *30*(1), 72–96. https://doi.org/10.1080/12460125.2020.1840705

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, *15*(3), 234–281. https://doi.org/10.1016/0022-2496(77)90033-5

Saaty, T. L. (1994). How to make a decision: The analytic hierarchy process. *Interfaces*, *24*(6), 19–43. https://doi.org/10.1287/inte.24.6.19

Saaty, T. L. (1996). The Analytic Network Process. In *Decision making with dependence and feedback*, Analytic hierarchy process series. USA: RWS Publications.

Sayeki, Y., & Vesper, K. H. (1973). Allocation of importance in a hierarchial goal structure. *Management Science*, *19*(6), 667–675. https://doi.org/10.1287/mnsc.19.6.667

von Nitzsch, R., & Weber, M. (1993). The effect of attribute ranges on weights in multiattribute utility measurements. *Management Science*, *39*(8), 937–943. https://doi.org/10.1287/mnsc.39.8.937

AUTHOR BIOGRAPHIES

Jafar Rezaei is an associate professor at the faculty of Technology, Policy and Management at TU Delft. His research interests include decision analysis and decision-making methodologies and their applications in (sustainable) logistics and supply chain management. He has developed several decision-making methods including best-worst method (BWM) which has become one of the mainstream methods in multiattribute decision-making. He is the editor-in-chief of *Journal of Supply Chain Management Science* and is in the editorial team of several scientific journals including *IEEE Transactions on Engineering Management*. He has organized several international events in decision-making including the EURO Working Group on MCDA and International Workshop on BWM. He has also delivered several keynotes on decision-making and supply chain management at some international conferences. He is interested in working on complex decision-making problems.

Alireza Arab is a PhD candidate in operations research at the Faculty of Management, University of Tehran. His main research interest areas include multiple-criteria decision-making (MCDM) and cognitive biases in MCDM. He has published in several scientific Operations Research journals, mainly on the topics of MCDM in different areas, including supply chain management and logistics, industry 4.0, risk management, and tourism management. He has been selected as the best student of the faculty of Management, University of Tehran in 2016.

Mohammadreza Mehregan is a full professor of Industrial Management at the Faculty of Management, University of Tehran. He obtained his PhD in Industrial management from Tarbiat Modarres University in 1994. His research interests include soft operations research, decision-making, mathematical modeling, and optimization. He serves as an Editorial Board and reviewer of many famous journals in the operations research field. He has been selected as the best teacher and researcher of the faculty and university several times. Professor Mehregan has published about 19 books and more than 500 papers in reputable academic journals and conferences.

How to cite this article: Rezaei, J., Arab, A., & Mehregan, M. (2021). Equalizing bias in eliciting attribute weights in multiattribute decision-making: experimental research. *Journal of Behavioral Decision Making*, 1–19. https://doi.org/10.1002/bdm.2262

APPENDIX A

The attributes and subattributes of the research problem

Attribute	Subattribute	References
Cost (C1)	Travel cost (C1–1)	(Celik et al., 2013; De Oña et al., 2013; Dirghayani & Sutanto, 2020; Eboli & Mazzulla, 2011; Mavi et al., 2018; Nassereddine & Eskandari, 2017)
Time (C2)	Travel time (C2–1)	(Deveci et al., 2019; Dirghayani & Sutanto, 2020; Errampalli et al., 2020; Hsu, 1999; Mavi et al., 2018; Nassereddine & Eskandari, 2017; Zak, 2011)
	Waiting time (C2–2)	(Celik et al., 2013; Dirghayani & Sutanto, 2020; Hsu, 1999; Nassereddine & Eskandari, 2017; Zak, 2011)

Attribute	Subattribute	References
	Reliability and punctuality of vehicles mode runs come on schedule to the destination (C2–3)	(Celik et al., 2013; De Oña et al., 2013; Eboli & Mazzulla, 2011; Errampalli et al., 2020; Kuo & Liang, 2012; Lee, 2018; Mavi et al., 2018; Zak, 2011)
Environment friendly (C3)	Pollution (C3–1)	(Celik et al., 2013; Deveci et al., 2019; Eboli & Mazzulla, 2011; Errampalli et al., 2020; Kuo & Liang, 2012; Lee, 2018; Mavi et al., 2018)
Comfort (C4)	The passenger density in the vehicle (C4–1)	(Celik et al., 2013; De Oña et al., 2013; Deveci et al., 2019; Dirghayani & Sutanto, 2020; Eboli & Mazzulla, 2011; Lee, 2018; Nassereddine & Eskandari, 2017; Wan et al., 2016; Zak, 2011)
	Ease of accessibility to vehicle stop station (C4–2)	(Celik et al., 2013; Cui et al., 2020; De Oña et al., 2013; Dirghayani & Sutanto, 2020; Errampalli et al., 2020; Hsu, 1999; Lee, 2018; Nassereddine & Eskandari, 2017)
	Air condition and other equipment in the vehicles (C4–3)	(Celik et al., 2013; Eboli & Mazzulla, 2011; De Oña et al., 2013; Deveci et al., 2019; Kuo & Liang, 2012; Lee, 2018; Zak, 2011)

REFERENCES FOR APPENDIX A

Celik, E., Bilisik, O. N., Erdogan, M., Gumus, A. T., & Baracli, H. (2013). An integrated novel interval type-2 fuzzy MCDM method to improve customer satisfaction in public transportation for Istanbul. *Transportation Research Part E: Logistics and Transportation Review*, *58*, 28–51.

Cui, B., Boisjoly, G., Miranda-Moreno, L., & El-Geneidy, A. (2020). Accessibility matters: Exploring the determinants of public transport mode share across income groups in Canadian cities. *Transportation Research Part D: Transport and Environment*, *80*, 102276.

De Oña, J., De Oña, R., Eboli, L., & Mazzulla, G. (2013). Perceived service quality in bus transit service: A structural equation approach. *Transport Policy*, *29*, 219–226.

Deveci, M., Öner, S. C., Canıtez, F., & Öner, M. (2019). Evaluation of service quality in public bus transportation using interval-valued intuitionistic fuzzy QFD methodology. *Research in Transportation Business & Management*, *33*, 100387.

Dirgahayani, P., & Sutanto, H. (2020). The effect of transport demand management policy on the intention to use public transport: A case in Bandung, Indonesia. *Case Studies on Transport Policy*, *8*(3), 1062–1072.

Eboli, L., & Mazzulla, G. (2011). A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view. *Transport Policy*, *18*(1), 172–181.

Errampalli, M., Patil, K. S., & Prasad, C. S. R. K. (2020). Evaluation of integration between public transportation modes by developing sustainability index for Indian cities. *Case Studies on Transport Policy*, *8*(1), 180–187.

Hsu, T. H. (1999). Public transport system project evaluation using the analytic hierarchy process: A fuzzy Delphi approach. *Transportation Planning and Technology*, *22*(4), 229–246.

Kuo, M. S., & Liang, G. S. (2012). A soft computing method of performance evaluation with MCDM based on interval-valued fuzzy numbers. *Applied Soft Computing*, *12*(1), 476–485.

Lee, D. J. (2018). A multi-criteria approach for prioritizing advanced public transport modes (APTM) considering urban types in Korea. *Transportation Research Part a: Policy and Practice*, *111*, 148–161.

Mavi, R. K., Zarbakhshnia, N., & Khazraei, A. (2018). Bus rapid transit (BRT): A simulation and multi criteria decision making (MCDM) approach. *Transport Policy*, *72*, 187–197.

Nassereddine, M., & Eskandari, H. (2017). An integrated MCDM approach to evaluate public transportation systems in Tehran. *Transportation Research Part a: Policy and Practice*, *106*, 427–439.

Wan, D., Kamga, C., Liu, J., Sugiura, A., & Beaton, E. B. (2016). Rider perception of a "light" bus rapid transit system—The new York City select bus service. *Transport Policy*, *49*, 41–55.

Zak, J. (2011). The methodology of multiple criteria decision making/aiding in public transportation. *Journal of Advanced Transportation*, *45*(1), 1–20.