




Is the best–worst method path dependent? Evidence from an empirical study

Jiri Mazurek¹ · Radomír Perzina¹ · Dominik Strzałka^{1,2}  · Bartosz Kowal² · Paweł Kuraś² · Barbora Petrů Puhrová³ · Robert Rajs⁴

Received: 31 March 2023 / Revised: 19 September 2023 / Accepted: 21 September 2023 /

Published online: 30 October 2023

© The Author(s) 2023

Abstract

The Best–Worst method (BWM) is one of the latest contributions to pairwise comparisons methods. As its name suggests, it is based on pairwise comparisons of all criteria (or possibly other objects, such as alternatives, sub-criteria, etc.) with respect to the best (most important) and the worst (least important) criterion. The main aim of this study is to investigate the path and scale dependency of the BWM. Up to now, it is unknown whether the weights of compared objects obtained by the method differ when the objects are compared first with the best object, and then with the worst, or vice versa. It is also unknown if the outcomes of the method differ when compared objects are presented in a different order, or when different scales are applied. Therefore, an experiment in a laboratory setting is performed with more than 800 respondents university undergraduates from two countries in which the respondents compare areas of randomly generated figures and the relative size of objects is then estimated via the linearized version of the BWM. Last but not least, the accuracy of the BWM is examined with respect to different comparison scales, including Saaty's nine-point linguistic scale, an integer scale from 1 to 9, and a continuous scale from 1 to infinity.

Keywords Pairwise comparisons · BWM · Best–Worst method · Path dependency · Scale dependency

Mathematics Subject Classification 90B50 · 91B06

✉ Dominik Strzałka
strzalka@prz.edu.pl

- ¹ Silesian University Opava, Opava, Czech Republic
- ² Rzeszów University of Technology, Rzeszow, Poland
- ³ Tomas Bata University in Zlin, Zlin, Czech Republic
- ⁴ Carpathian State College in Krosno, Krosno, Poland

1 Introduction

Path dependency refers to the impact of a method's problem solving path on its outcome (Hämäläinen and Lahtinen 2016). In general, a path is a sequence of steps undertaken in a problem solving process and might include, for example, an order in which the different parts of the model are specified and solved, or a way in which data or preferences are collected and processed (Hämäläinen and Lahtinen 2016). In Hämäläinen and Lahtinen (2016), seven interacting origins of path dependence are identified: systemic origins, learning, procedure, behavior, motivation, uncertainty, and external environment.

In general, the path dependence of a model or method, for instance, a multiple-criteria decision-making method, is perceived as a negative phenomenon: path dependency of a problem solving method means that, when some steps of the method are performed in a different order, the outcome (usually the best solution) might also be different, which is, of course, an undesirable result.

Interestingly, studies on this topic in the field of operations research are still rare in the literature. The possibility that different 'valid' modeling paths lead to different outcomes was acknowledged by Landry et al. (1983) in the 1980s, but the topic received little interest in the operations research community. Three decades later, path dependence attracted the attention of the studies (Hämäläinen and Lahtinen 2016; Lahtinen and Hämäläinen 2016; Lahtinen et al. 2017), where the latter examined path dependence in the even-swaps method.

The influence of a selected comparison scale on outcomes of multiple criteria decision-making methods, i.e. the scale dependence, has been examined much more. In particular, the problem of a scale in the analytic hierarchy process (AHP) has been studied extensively in recent decades, see e.g. Franek and Kresta (2014), Harker and Vargas (1987), Ishizaka et al. (2010), Leskinen (2008), Ma and Zheng (1991), Poyhonen and Hämäläinen (2001), Salo and Hämäläinen (1997), Setyono and Sarno (2018). For the AHP, Saaty proposed the use of the so-called 'fundamental scale' from 1 to 9 (with reciprocal values) based on psychological arguments, see Saaty (1977). Later, many other scales, such as exponential, logarithmic, interval and so on, were proposed for the AHP, but no scale appeared to be superior to the others in general. In the BWM, the same scale from 1 to 9 is applied, though in the original paper (Rezaei 2015) Rezaei notes that other scales can be used as well. However, the effect of the scale on the BWM's outcomes is currently unknown.

The Best–Worst method (BWM) proposed by Rezaei in 2015 (Rezaei 2015) is one of the most recent contributions to the decision-making methods based on pairwise comparisons, and immediately after its introduction gained significant popularity among researchers and has been applied in many areas of human activity, such as waste management, tourism, sustainability or biochemistry (Abadia et al. 2018; Ahmadi et al. 2017; Chang et al. 2019; Gupta and Barua 2017; Mi et al. 2019; Rezaei et al. 2016; Thurstone 1927).

In the BWM, a decision maker compares each criterion (or possibly some other object, alternative, sub-criterion, etc.) only with the best (most important) criterion and the worst (least important) criterion, and then the weights of all criteria are determined by the solution of a (non)linear programming problem. The appeal of the BWM lies

in its simplicity and the smaller number of pairwise comparisons necessary to be performed when compared to the AHP, however, as shown in Mazurek et al. (2021), its robustness is weaker than for the geometric mean method and the eigenvalue method. The problem of whether the BWM is path independent, or scale independent (robust to changes in which comparisons are performed, or robust to changes in scales by which these comparisons are performed) is currently unresolved, as no study on the topic has been published in the literature so far.

Therefore, the primary objective of this paper was to examine the path and scale dependency of the linear version of the BWM, which does not suffer the problem of possible multiple solutions as its non-linear version, via an experiment in which more than 800 respondents from two countries (Czechia and Poland) pairwise compared by estimation (without measuring with a ruler, or calculating) areas of six geometric objects presented in different orders and via different comparison scales. Altogether, five distinct questionnaire forms were distributed among the respondents. Subsequently, these preferences served as an input for the BWM and the output consisted of relative sizes of the compared objects for each form.

Afterwards, the path and scale differences among forms were tested by ANOVA and MANOVA, the multivariate extension of ANOVA, see Allen et al. (2018), Barker and Barker (1984), Brown (1998), Olson (1976), Weinfurt (1995), Zaointz (2022). Post-hoc analysis of the results was performed as well. The secondary objective of our study was to examine the accuracy of the BWM with respect to different scales, including Saaty's nine-point linguistic scale, integer scale and continuous scale.

The study falls, at least partially, into behavioral operational research (BOR) category, that is the research field dedicated to understanding how the behavior of human actors influences their decisions, see e.g. Brocklesby (2016), Franco et al. (2021), Hämäläinen et al. (2013), Kunc et al. (2016). In particular, the BOR focuses on cognitive biases, which are systematic errors in human judgements. That's why the problem of a possible cognitive bias in the presented experiment is discussed in a separate section (Discussion see Sect. 5) as well.

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

The paper is organized as follows: Sect. 2 provides an introduction to the Best–Worst method, the experiment is described in Sect. 3, Sect. 4 summarizes the experiment results, in Sect. 5 a brief Discussion is provided and the Conclusions (Sect. 6) close the article.

2 The best–worst method

In the Best–Worst method, see Rezaei (2015, 2016), each criterion is pairwise compared only with the best criterion and the worst criterion.

The Best–Worst method proceeds as follows (Rezaei 2015):

Step 1. A set of criteria is determined.

Step 2. The decision maker identifies the best (most desirable, most important) criterion and the worst (least desirable, least important) criterion.

Step 3. Preferences of the best criterion with respect to all other criteria are determined on the scale from 1 (equal preference) to 9 (absolute preference).

Step 4. Preferences of all other criteria with respect to the worst criterion are determined on the scale from 1 to 9.

Step 5. Optimal weights of all criteria are found by solving a corresponding non-linear programming problem.

Let c_{Bj} denote the preference of the best criterion (B) over the criterion (j), whereas let c_{iW} denote the preference of the criterion (i) over the worst criterion (W). Let w_B and w_W be the weights of the best and worst criterion, respectively. The goal is to find the vector of criteria weights (a priority vector) $w = (w_1, w_2, \dots, w_n)$.

The priority vector is found as a solution of the following optimization problem (Rezaei 2015):

$$\min \left(\max_j \left\{ \left| \frac{w_B}{w_j} - c_{Bj} \right|, \left| \frac{w_j}{w_W} - c_{jW} \right| \right\} \right) \quad (1)$$

s.t.

$$\sum_{j=1}^n w_j = 1 \quad (2)$$

$$w_j \geq 0, \quad \forall j = 1, \dots, n. \quad (3)$$

The problem can equivalently be stated as follows:

$$\min \xi \quad (4)$$

s.t.

$$\left| \frac{w_B}{w_j} - c_{Bj} \right| \leq \xi, \quad \forall j = 1, \dots, n, \quad (5)$$

$$\left| \frac{w_j}{w_W} - c_{jW} \right| \leq \xi, \quad \forall j = 1, \dots, n, \quad (6)$$

$$\sum_{j=1}^n w_j = 1, \quad (7)$$

$$w_j \geq 0, \quad \forall j = 1, \dots, n. \quad (8)$$

Further on, it is assumed that for all j the following inequalities hold:

$$C_{BW} \geq C_{jW} \geq 1; \quad C_{BW} \geq C_{Bj} \geq 1. \quad (9)$$

A linear version of the BWM was introduced by Brunelli and Rezaei (2019), Rezaei (2016), where the letter ‘ L ’ denotes linear:

$$\min \xi_L \quad (10)$$

s.t.

$$|w_B - c_{Bj}w_j| \leq \xi_L, \quad \forall j = 1, \dots, n, \quad (11)$$

$$|w_j - c_{jW}w_W| \leq \xi_L, \quad \forall j = 1, \dots, n, \quad (12)$$

$$\sum_{j=1}^n w_j = 1, \quad (13)$$

$$w_j \geq 0, \quad \forall j = 1, \dots, n. \quad (14)$$

The solution of the model above is denoted as w^* and the corresponding value of ξ_L^* can be considered the degree of inconsistency of preferences. Notice that the solution to the linear version of the BWM differs from the solution to the non-linear version in general (Beemsterboer et al. 2018).

When comparing n objects pairwise, the analytic hierarchy process (AHP), which is arguably the most popular pairwise comparisons method, requires $n(n - 1)/2$ comparisons to be made. The BWM requires only comparisons with the best and worst object, and the reduced number of comparisons amounts to $2n - 3$. This reduction might be very important when dealing with a large number of objects to compare.

3 The experiment, research hypotheses and the data

3.1 The description of the experiment

For the investigation of the path and scale dependency of the Best–Worst method, the following experiment was designed.

Participants of the research respondents were university undergraduate students aged 19–22. The study was anonymous and the authors had no access to information that could identify individual participants. Questionnaires were distributed to respondents in classrooms in groups from 10 to 25 and respondents consented to participate in the study verbally at the beginning of the experiment.

Then, respondents were asked to pairwise compare by estimation (measurements with a ruler, or calculations, were not allowed), the areas of six geometric objects, see Fig. 1 and a sample (filled) questionnaire in Appendix A. Respondents answered simple questions of the following type: How many times is the area of the triangle greater than the area of the circle?, and the answer was written into a predefined blank spot.

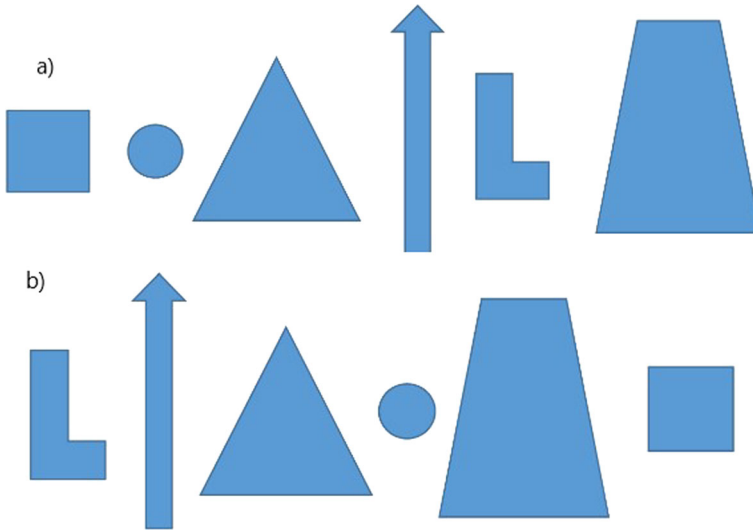


Fig. 1 The order of compared objects in questionnaires: **a** A, B, D, and E; **b** C

Questionnaires were divided into 5 different forms: A, B, C, D, and E (see the description below), with the same figures of exactly the same size, but with different orders of the presented figures or a different comparison scale. The questionnaires were distributed in a printed (paper) form.

Altogether, 846 respondents from four universities in Czechia and Poland; Silesian University in Opava (CZ), Tomas Bata University in Zlin (CZ), Rzeszów University of Technology (PL) and Carpathian State College in Krosno (PL), took part in the experiment.

The respondents consisted of 470 men (55.6%) and 376 women (44.4%). Each respondent filled exactly one questionnaire (one form). The numbers of respondents and their gender are reported in Table 1. The ratio of men and women for each questionnaire was roughly the same to minimize possible gender differences in the perception of the figures. Questionnaires were printed in respondents' native language, that is Czech or Polish.

Table 1 Forms and respondents' numbers

Form	Count	Men (%)
A	173	57
B	156	56
C	197	53
D	148	55
E	172	57
Total	846	55.6

The figure with the largest area (Best) was the trapezoid, the figure with the smallest area (Worst) was the circle. figures' areas were generated randomly. Further on, the two orderings of figures shown in Figs. 1a,b) and used for the experiment were also generated randomly (as a permutation of six objects).

Form A: See Fig. 1a. All objects were compared with the Best object, and then with the Worst object via scale $[1, \infty[$.

Form B: See Fig. 1a. All objects were compared with the Worst object, and then with the Best object via scale $[1, \infty[$.

Form C: See Fig. 1b. All objects were compared with the Best object, and then with the Worst object via scale $[1, \infty[$, but the order of compared objects was different than for other forms.

Form D: See Fig. 1a. All objects were compared with the Best object, and then with the Worst object via Saaty's scale from 1 to 9 (with reciprocals).

Form E: See Fig. 1a. All objects were compared with the Best object, and then with the Worst object via Saaty's linguistic scale, see Saaty (1977, 1980).

The weights of objects (their relative sizes) from questionnaires A-E were derived via a linearized version (Eqs. (10)–(14)) of the BWM. The experiment results are provided in the next section.

3.2 The research hypotheses

In order to investigate path and scale dependence of the BWM, the following null hypotheses were formulated and tested (the letter μ denotes the mean value operator).

$$H_{01} : \mu(w_i^A) = \mu(w_i^B) = \mu(w_i^C), \quad i \in \{1, \dots, 6\}.$$

The hypothesis H_{01} deals with the path dependency of the Best–Worst method, which, in the experiment's setting, can be realized in two different ways: firstly, the order of comparisons with the Best and Worst object can be reversed, and, secondly, the order of comparisons of individual pairs can change, see Fig. 1. If H_{01} holds for all indices i , then the mean values of weights of all objects are the same in cases of questionnaires A, B and C (which differ in paths, but not scales), which means the BWM is not path dependent. Otherwise, the null hypothesis is rejected and the BWM is path dependent.

Further on, both cases of 'path changes' mentioned above are examined separately, hence, two additional null hypotheses were formulated:

$$H_{01a} : \mu(w_i^A) = \mu(w_i^B), \quad i \in \{1, \dots, 6\}.$$

$$H_{01b} : \mu(w_i^A) = \mu(w_i^C), \quad i \in \{1, \dots, 6\}.$$

The hypothesis H_{01a} states that results of the BWM do not depend on the order of comparisons of all objects with respect to the Best and the Worst object, respectively.

The hypothesis H_{01b} states that results of the BWM do not depend on the order in which all objects are mutually compared pairwise.

The next hypothesis H_{02} deals with the scale dependency of the BestWorst method:

$$H_{02} : \mu(w_i^A) = \mu(w_i^D) = \mu(w_i^E), \quad i \in \{1, \dots, 6\}.$$

If H_{02} holds for all indices i , then values of weights of all objects are the same in cases of questionnaires A, D and E (which differ in scales but have the same path), which means the BWM is scale invariant. Otherwise, the null hypothesis is rejected and the BWM is scale dependent.

As with the hypothesis H_{01} , one particular subcase of H_{02} is examined as well, namely difference of the BWM results for integer Saaty's scale and real scale from 1 to in nity, that is questionnaires A and D:

$$H_{02a} : \mu(w_i^A) = \mu(w_i^D), \quad i \in \{1, \dots, 6\}.$$

Finally, the last hypothesis H_{03} deals with the accuracy of respondents' comparisons with respect to the three comparison scales (see forms A, D and E), which is both interesting and important from a practical point of view.

Here, the accuracy is estimated via the mean relative error, where the actual relative size of all six objects is denoted as $w^* = (w_1^*, \dots, w_6^*)$. Formally, the mean relative error of a respondent j filling the form q is given as follows:

$$d_j^q = \frac{1}{6} \sum_{i=1}^6 \frac{|w_{i,j}^q - w_i^*|}{w_i^*} \quad (15)$$

The following null hypothesis states that respondents were, on average, equally accurate in their judgments transformed by the BWM into the relative sizes of the compared objects for all three scales.

$$H_{03} : \mu(d_j^A) = \mu(d_j^D) = \mu(d_j^E).$$

3.3 The data

To ensure the data quality, respondents' responses were assessed and deficient questionnaires were discarded on the following grounds:

- (i) A questionnaire was incomplete.
- (ii) A questionnaire did not conform to instructions for its filling (most often respondents used wrong scale for comparisons).
- (iii) A questionnaire included outliers.

Outliers' identification was performed via SPSS (Tukey's test) and via Gretl (Mahalanobis distance).

Altogether, approximately 13% of the questionnaires were removed from the dataset.

3.4 The hypotheses testing—MANOVA

In the first two hypotheses, not one independent variable, but six independent variables (areas of six geometrical objects) are assessed at once. Therefore, these hypotheses were tested via one-way multivariate analysis of variance (MANOVA). The third hypothesis included only one independent variable, hence it was tested by one-way ANOVA.

According to (Warne 2014; Zientek and Thompson 2009), MANOVA is a member of the General Linear Model class, a family of statistical procedures often used to quantify the strength between variables. MANOVA extends the capabilities of analysis of variance (ANOVA) by assessing multiple dependent variables simultaneously. This provides several advantages: when the dependent variables are correlated, MANOVA can identify effects that are smaller than those that regular ANOVA can detect. Further on, MANOVA can assess patterns between multiple dependent variables, which ANOVA cannot. Additionally, MANOVA limits the joint error rate. When a series of ANOVA tests is performed, the joint probability of rejecting a true null hypothesis increases with each additional test (thus Bonferroni or another correction is necessary), but in MANOVA the error rate equals the significance level.

Similarly to ANOVA, MANOVA has several assumptions (Anderson et al. 1996; Zientek and Thompson 2009):

- Observations are randomly and independently sampled from the population.
- Each dependent variable is measured at the interval or ratio level.
- An independent variable consists of two or more categorical (independent) groups.
- Dependent variables are multivariate normally distributed.
- The population covariance matrices of each group are equal (homogeneity of variance–covariance matrices).

Other sources add the absence of outliers and the absence of multicollinearity of dependent variables as well, see e.g. Barker and Barker (1984), Finch (2005).

MANOVA provides four statistics for hypothesis testing: Pillai's trace, Wilk's lambda, Hotelling's trace and Roy's greatest root. In the case of two groups (see hypotheses H_{01a} , H_{01b} and H_{02a}), all the statistics are equivalent and the test reduces to Hotelling's T-square.

It should be noted that though MANOVA is a very useful statistical tool, it has also its limitations. Discussion continues over the merits of each statistic mentioned above, see e.g. Weinfurt (1995), and about violation of MANOVA's assumptions on its preformation. In particular, according to Finch (2005), MANOVA is robust against departures from multivariate normality especially when the number of data points is large. The study (Knief and Forstmeier 2021) found that Gaussian models (such as ANOVA) are remarkably robust to non-normality over a wide range of conditions, meaning that P -values remain fairly reliable except for data with influential outliers. Also, it is argued in Allen et al. (2018); Olson (1976) that MANOVA is robust against violations of homogeneity of variance–covariance matrices assumption. When there

is a violation of the equality of variances, Pillai's trace is the most suitable characteristic for MANOVA, as it is highly robust to many violations of the assumptions of MANOVA, see e.g. Allen et al. (2018), Finch (2005), Olson (1976).

4 Results of the experiment

After the data were gathered from the participants of the experiment, the weights of all six geometric objects corresponding to their relative sizes were calculated by the linear version of the BWM (Eqs. (10)–(14)) for each respondent/questionnaire. Afterwards, the data were checked for outliers, which were removed from the dataset.

4.1 Descriptive statistics

Next, the descriptive statistics of all questionnaire forms A-E were performed separately for the sake of comparison. The relative sizes of the compared geometric objects in the form of the weights w_i^q , where $i \in \{1, \dots, 6\}$, $q \in \{A, B, C, D, E\}$ are summarized in Table 2. The last row of Table 2 contains actual relative sizes of all six objects, see also Fig. 2. As can be seen, objects' relative sizes derived from questionnaire forms A-E were close to each other (with the exception of Trapezoid in the questionnaire E) and to the actual relative sizes of objects. Interestingly, respondents of forms A-D underestimated the area of the Trapezoid on average and, simultaneously, overestimated the area of the Square, Arrow, L-Shape and Circle.

In many real-world problems, the ranking of compared objects is more important to a decision maker than precise values of a priority vector. The ranking of all six geometric figures with respect to questionnaire form is shown in Table 3. The correct ranking was obtained from forms A, B and D. Form E contained one discordant pair (Arrow-Square), while form C contained two discordant pairs (L-Square and L-Arrow). The difference between the ranking A (or B and D) and C can be expressed via Kendall's rank correlation coefficient as $\tau(A, C) = 0.733$. Other pairwise Kendall's rank correlation coefficients were even higher, hence the rankings obtained from different questionnaire forms were highly correlated (similar).

4.2 Path dependence of the BWM

To test the path dependence of the BWM (the null hypotheses H_{01} , H_{01a} and H_{01b}), the data containing six dependent variables weights of six geometric objects corresponding to their relative sizes (the variables are denoted simply as Trapezoid, Square, Triangle, Arrow, L and Circle) and one independent variable the questionnaire form (A, B and C), was prepared.

Before the null hypotheses were tested via SPSS, the assumptions of MANOVA were checked:

- The data were properly randomly and independently sampled from the population.
- The dependent variables were ratios.
- The independent variable consisted of three (two) independent groups.

Table 2 Experiment results: the mean weights (relative sizes) of all geometric objects for all questionnaire forms with the standard deviation in parentheses and the number (N) of questionnaires (after the data quality check)

Form	N	Trapezoid	Square	Triangle	Arrow	L	Circle
A	150	0.388 (0.032)	0.118 (0.018)	0.237 (0.026)	0.113 (0.021)	0.103 (0.019)	0.040 (0.007)
B	134	0.397 (0.037)	0.116 (0.020)	0.226 (0.036)	0.115 (0.025)	0.103 (0.019)	0.043 (0.010)
C	185	0.387 (0.041)	0.110 (0.024)	0.249 (0.037)	0.101 (0.031)	0.111 (0.027)	0.041 (0.009)
D	130	0.389 (0.033)	0.119 (0.022)	0.229 (0.026)	0.117 (0.028)	0.103 (0.022)	0.042 (0.006)
E	140	0.457 (0.040)	0.097 (0.019)	0.204 (0.048)	0.107 (0.021)	0.095 (0.020)	0.041 (0.005)
Precise values		0.4208	0.1099	0.2278	0.1058	0.0974	0.0382

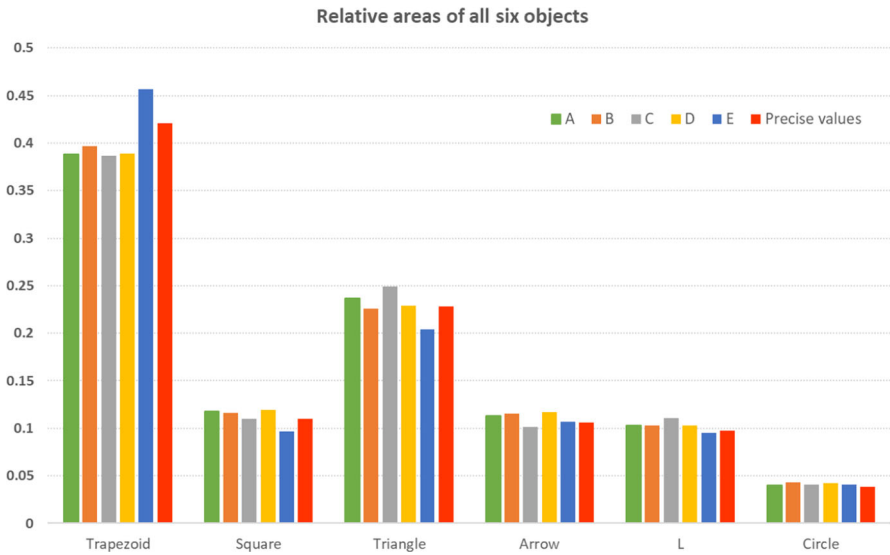


Fig. 2 The relative sizes of all objects with respect to all questionnaire forms (A)–(E)

Table 3 Experiment results: the ranking of figures’ sizes in descending order for all questionnaire forms

Rank/form	A	B	C	D	E
1	Trapezoid	Trapezoid	Trapezoid	Trapezoid	Trapezoid
2	Triangle	Triangle	Triangle	Triangle	Triangle
3	Square	Square	L	Square	Arrow
4	Arrow	Arrow	Square	Arrow	Square
5	L	L	Arrow	L	L
6	Circle	Circle	Circle	Circle	Circle

- The correlation matrix revealed no significant multicollinearity of the dependent variables (no correlation coefficient exceeded the absolute value of 0.40).
- Dependent variables (for all questionnaire forms) were tested for normality in Gretl via Shapiro–Wilk test. The result in the form of *p*-values shown in Table 4. As can be seen, normality couldn’t be rejected at 0.001 level for all variables.
- The homogeneity of variance–covariance matrices was tested via Levene’s test in SPSS. This property was violated in the case of Triangle, Arrow, L-shape and Circle at 0.001 level.

Though the last property was not satisfied for four objects, MANOVA is robust against the violation of the homogeneity variance–covariance matrices as pointed out in the Sect. 3.4.

Therefore, the hypothesis H_{01} was tested by MANOVA.

Table 4 *p*-values for the null hypothesis that the data follow the normal distribution via Shapiro–Wilk test in Gretl

Form/object	Trapezoid	Square	Triangle	Arrow	L	Circle
A	0.042	0.119	0.003	0.009	0.150	0.001
B	0.538	0.006	0.009	0.710	0.052	0.004
C	0.408	0.740	0.002	0.042	0.002	0.024
D	0.001	0.812	0.001	0.025	0.005	0.015
E	0.321	0.002	0.001	0.034	0.002	0.128

MANOVA SPSS' output is shown in Table 5. According to all four test characteristics: Pillai's trace, Wilks' Lambda, Hotelling's Trace and Roy's.

Largest Root, the hypothesis H_{01} was rejected at the 0.001 significance level. Therefore, it can be concluded that the weights corresponding to the relative sizes of six geometric objects obtained from questionnaires A, B and C differed significantly, hence the BWM was found to be path dependent.

Table 5 Multivariate tests for H_{01} , SPSS output

Effect		Value	F	Hypothesis df	Error df	Sig
<i>Multivariate test^a</i>						
Intercept	Pillai's Trace	1.000	1,220,812.543 ^b	5.000	462.000	0.000
	Wilks' Lambda	0.000	1,220,812.543 ^b	5.000	462.000	0.000
	Hotelling's Trace	13,212.26	1,220,812.543 ^b	5.000	462.000	0.000
	Roy's Largest Root	13,212.26	1,220,812.543 ^b	5.000	462.000	0.000
Form	Pillai's Trace	0.162	8.177	10.000	926.000	< 0.001
	Wilks' Lambda	0.842	8.296 ^b	10.000	924.000	< 0.001
	Hotelling's Trace	0.183	8.414	10.000	922.000	< 0.001
	Roy's Largest Root	0.148	13.710 ^c	5.000	463.000	< 0.001

^aDesign: Intercept + Form

^bExact statistic

^cThe Statistic is an upper bound on F that yields a lower bound on the significance level

Since the hypothesis H_{01} was rejected, the MANOVA was followed by post-hoc analysis to determine the source of the differences behind a rejection of a null hypothesis (Weinfurt 1995). Separate ANOVA tests of Between-Subjects Effects revealed that the most statistically significant differences occurred for the Triangle, Arrow and L-shape. Consequently, pairwise Fisher's Least Significant Difference (LSD) tests were performed to find out a statistical significance of differences in area estimations of the compared objects with respect to the three questionnaire forms. The cases with the statistical significance lower than 0.001 include the Square in the forms A-C, Triangle in forms B-C, the Arrow in forms A-C and B-C, and finally the L-Shape in B-C. These are the greatest differences among forms and thereby the main sources of the rejection of the hypothesis H_{01} .

Next, the hypotheses H_{01a} and H_{01b} dealing with different forms of path dependency were tested by MANOVA via SPSS as well.

In the case of the hypothesis H_{01a} (forms A and B), the homogeneity of variance—covariance matrices (tested via Levene's test) was satisfied for all objects with the only exception of Circle. The SPSS output of the hypothesis test is shown in Table 6. As can be seen, the hypothesis was rejected at 0.001 level. This means that the BWM results depended on the order of comparisons of all objects with the Best and the Worst object, respectively.

In the case of the hypothesis H_{01b} (forms A and C), the homogeneity of variance—covariance matrices (tested via Levene's test) was satisfied for Trapezoid, Square and

Table 6 Multivariate tests for H_{01a} , SPSS output

Effect		Value	F	Hypothesis df	Error df	Sig
<i>Multivariate test^a</i>						
Intercept	Pillai's Trace	1.000	861,320.236 ^b	5.000	278.000	0.000
	Wilks' Lambda	0.000	861,320.236 ^b	5.000	278.000	0.000
	Hotelling's Trace	15,491.371	861,320.236 ^b	5.000	278.000	0.000
	Roy's Largest Root	15,491.371	861,320.236 ^b	5.000	278.000	0.000
Form	Pillai's Trace	0.074	4.473 ^b	5.000	278.000	< 0.001
	Wilks' Lambda	0.926	4.473 ^b	5.000	278.000	< 0.001
	Hotelling's Trace	0.080	4.473 ^b	5.000	278.000	< 0.001
	Roy's Largest Root	0.080	4.473 ^b	5.000	278.000	< 0.001

^aDesign: Intercept + Form

^bExact statistic

Table 7 Multivariate tests for H_{01b} , SPSS output

Effect		Value	F	Hypothesis df	Error df	Sig
<i>Multivariate test^a</i>						
Intercept	Pillai's Trace	1.000	955,792.521 ^b	5.000	329.000	0.000
	Wilks' Lambda	0.000	955,792.521 ^b	5.000	329.000	0.000
	Hotelling's Trace	14,525.722	955,792.521 ^b	5.000	329.000	0.000
	Roy's Largest Root	14,525.722	955,792.521 ^b	5.000	329.000	0.000
Form	Pillai's Trace	0.114	8.491 ^b	5.000	329.000	< 0.001
	Wilks' Lambda	0.886	8.491 ^b	5.000	329.000	< 0.001
	Hotelling's Trace	0.129	8.491 ^b	5.000	329.000	< 0.001
	Roy's Largest Root	0.129	8.491 ^b	5.000	329.000	< 0.001

^aDesign: Intercept + Form^bExact statistic

Circle, and violated for the rest. The SPSS output of the hypothesis test is shown in Table 7. As can be seen, the hypothesis was rejected at 0.001 level. This means that the BWM results depended on the order of mutual pairwise comparisons of all objects.

4.3 Scale dependence of the BWM

Originally, the most suitable scale proposed for the BWM was Saaty's (numerical) scale from 1 to 9 (Rezaei 2015), nevertheless, the author mentioned that other scales can be used as well. Therefore, in this study, Saaty's linguistic scale (see form E) and a continuous scale (form A) are considered as well and compared with the integer 1–9 scale (form D). A general discussion on the type of scale that can or cannot be used in pairwise comparisons can be found in Koczkodaj et al. (2020), Mazurek (2023). After respondents provided their answers, Saaty's linguistic scale was transformed (for obvious computational reasons) to the integer 9 point scale per Saaty's mutual correspondence ('equal size' = 1, 'equally to moderately larger' = 2, etc.), though this correspondence was criticized in the past (however, this correspondence is still widely used in practice).

To test scale dependence of the BWM, that is the null hypothesis H_{02} , the data file containing six dependent variables weights of six geometric objects corresponding to their relative sizes (Trapezoid, Square, Triangle, Arrow, L and Circle) and one

independent variable the questionnaire form (A, D and E), was prepared in the same way as in the previous section.

Before the testing of the hypothesis H_{02} via MANOVA in SPSS, MANOVA assumptions were checked in the same way as in the previous section for hypothesis H_{01} . The data satisfied all assumptions with the only one exception regarding homogeneity of variance–covariance matrices, which was violated for Triangle.

The result of the test of the hypothesis H_{02} via MANOVA is reported in Table 8. According to all four test characteristics: Pillai's trace, Wilks' Lambda, Hotelling's Trace and Roy's Largest Root, the hypothesis was rejected at the 0.001 significance level. Therefore, it can be concluded that the weights corresponding to the relative sizes of six geometric objects obtained from questionnaires A, D and E differed significantly, hence the BWM was found to be scale dependent.

Post-hoc analysis revealed that the most significant differences among the questionnaire forms occurred for all geometric figures with the exception of the Circle. Fisher's Least Significant Difference (LSD) pairwise tests found that the relative sizes of the Trapezoid, Square, Triangle, Arrow and L-Shape were all statistically different at the 0.001 level for the forms A-E and D-E. Since form E included a linguistic scale, it can be concluded that estimates with this scale differed from the numerical scales in forms A and D.

Table 8 Multivariate tests for H_{02} , SPSS output

Effect		Value	F	Hypothesis df	Error df	Sig
<i>Multivariate Test^a</i>						
Multivariate Test ^a	Pillai's Trace	1.000	1,894,777.072 ^b	5.000	413.000	0.000
	Wilks' Lambda	0.000	1,894,777.072 ^b	5.000	413.000	0.000
	Hotelling's Trace	22,939.190	1,894,777.072 ^b	5.000	413.000	0.000
	Roy's Largest Root	22,939.190	1,894,777.072 ^b	5.000	413.000	0.000
Form	Pillai's Trace	0.516	28.828	10.000	413.000	< 0.001
	Wilks' Lambda	0.494	34.865 ^b	10.000	413.000	< 0.001
	Hotelling's Trace	1.000	41.206	10.000	413.000	< 0.001
	Roy's Largest Root	0.997	80.933 ^c	5.000	413.000	< 0.001

^aDesign: Intercept + Form

^bExact statistic

^cThe Statistic is an upper bound on F that yields a lower bound on the significance level

Table 9 Multivariate tests for H_{02a} , SPSS output

Effect		Value	F	Hypothesis df	Error df	Sig
<i>Multivariate test^a</i>						
Intercept	Pillai's Trace	1.000	1,095,612.338 ^b	5.000	274.000	0.000
	Wilks' Lambda	0.000	1,095,612.338 ^b	5.000	274.000	0.000
	Hotelling's Trace	19,992.926	1,095,612.338 ^b	5.000	274.000	0.000
	Roy's Largest Root	19,992.926	1,095,612.338 ^b	5.000	274.000	0.000
Form	Pillai's Trace	0.039	2.202 ^b	5.000	274.000	0.054
	Wilks' Lambda	0.961	2.202 ^b	5.000	274.000	0.054
	Hotelling's Trace	0.040	2.202 ^b	5.000	274.000	0.054
	Roy's Largest Root	0.040	2.202 ^b	5.000	274.000	0.054

^aDesign: Intercept + Form

^bExact statistic

Next, the null hypothesis H_{02a} dealing with 1–9 scale (form D) and $[1, \infty)$ scale (form A) was tested via MANOVA as well. All MANOVA assumptions were checked and found satisfied including homogeneity of variance–covariance matrices for all six objects. The result of MANOVA is reported in Table 9. All four MANOVA statistics showed $p = 0.054$, which means the H_{02a} hypothesis couldn't be rejected at 0.05 level. Therefore, it can be concluded that there is no statistically significant difference in the use of both scales.

4.4 Accuracy of the BWM

In addition to the investigation of the path and scale dependency of the BWM, the accuracy of respondents' estimations was examined for each form A–E via relation (15). The results are shown in Table 10 and Fig. 3.

Respondents of the questionnaire form A were the most precise in their estimations (with the mean relative error of 13.1%), while respondents of the forms C (with the

Table 10 Average respondents' mean relative error ($(\mu(d^q_j))$) in % for all questionnaire forms

Form	A	B	C	D	E
Mean relative error	13.1	15.4	17.4	15.0	16.0

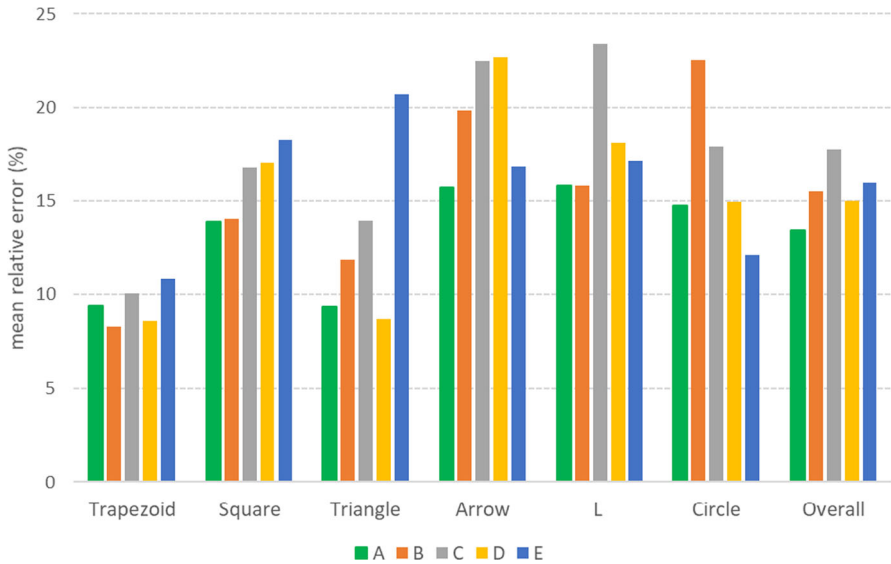


Fig. 3 Mean relative error in estimations of figures’ areas for all questionnaire forms and all figures

mean relative error of 17.4%) were least precise in their judgments. As for geometric figures, respondents were most accurate in the estimation of the relative size of Trapezoid (the mean relative error of 9.4%) and least accurate for Arrow (the mean relative error of 19.5%), probably due to its complex shape.

Next, the accuracy of the BWM with respect to three different scales continuous (form A), integer (form D) and linguistic (form E) was evaluated.

The null hypothesis H_{03} was tested via one-way ANOVA, where the independent variable was the form (i.e. scale) and the dependent variable was the mean relative error, see relation (15). The dataset contained only one outlier, which was removed.

Before the testing assumptions of ANOVA were checked. The normality of the data couldn’t be rejected at 0.001 level. Multicollinearity of the data was not detected (correlation coefficients were lower than 0.10 in the absolute value). The equality of variances couldn’t be rejected at 0.01 level.

The results of ANOVA are shown in Table 11. The p -value was $3.3 \cdot 10^{-5}$, which means the hypothesis H_{03} can be rejected at both 0.05 and 0.01 level. Therefore, comparisons’ scale had statistically significant impact on accuracy of comparisons.

The lowest mean relative error in comparisons occurred in the case of form A, that is continuous scale from 1 to infinity. It is a rather expected result since allowing

Table 11 ANOVA results for the hypothesis H_{03}

Form	$\mu(d^q_j)$	$\sigma^2(dqj)$	ANOVA
A	0.131	0.0030	$F = 10.59$ $p = 3.26 \cdot 10^{-5}$
D	0.150	0.0036	
E	0.160	0.0021	

decision makers to use continuous scale rather than integer scale may contribute to more accurate judgments in cases when, for example, a decision maker is not sure whether an object M is 2 times or 3 times more important (more preferred, bigger, etc.) than an object N. Then, a value between 2 and 3 can be assigned. The integer 1–9 scale (form D) and linguistic scale (form E) were found almost identically accurate.

5 Discussion

The results of the experiment described in the previous section suggest that the Best–Worst method (BWM) is both path and scale dependent. Reasons behind this outcome might be twofold.

Firstly, one possible cause is human cognitive bias. It is well-known that human thinking is susceptible to many systematic errors such as anchoring bias, attentional bias, attribution bias, framing effect, confirmation bias, recency bias, response bias and many others. Further on, there are many studies on human perception regarding geometric shapes and their areas, see e.g. Krider et al. (2001) for a review. When comparing areas of geometric objects, two main factors are shape and size. Researchers found, for example, that triangles have been generally found to be perceived larger than circles and squares (of the same area), or that elongated figures were perceived larger (Krider et al. 2001). Also, areas of complex shapes were found harder to be accurately estimated. As can be seen from Fig. 3, in our experiment the area of the most complex shape, Arrow, was estimated with the greatest error indeed.

Latimer et al. (2000) investigated performance in the perception of simple geometric forms placed at the top and to the right of the visual field rather than top-left, bottom-right or bottom-left with the result that figures at top-right were processed faster than others, and this ‘top-right’ bias was statistically significant. Therefore, the placement of figures to be compared matters. At last, but not least, a loss of focus might have occurred to respondents: they might focus on just few comparisons more than on the last ones.

Secondly, the path and scale dependence of the BWM might be a consequence of the transformation of respondents’ judgments into objects’ weights via linear model (Eqs. (10)–(14)).

The design of the conducted experiment does not allow to make conclusions about the cause of the dependence since its objective was different. However, it is likely that both human bias and transformation of judgments played a role. It should be noted that if a human cognitive bias is the root of BWM’s path and scale dependency, then this bias will be very likely present also in other pairwise comparisons methods of a similar design such as the analytic hierarchy process (AHP).

It should be noted that the experiment described in previous sections was based on one criterion only: the area of figures, while the BWM is generally a multiple-criteria decision aiding method. However, one criterion is sufficient for the investigation of path and scale dependency of the BWM and enables to draw conclusions without the necessity of tackling complexity of multiple criteria design. Nevertheless, in the multiple-criteria version of the BWM the so called “splitting bias” might be present (when one criterion is split into two different criteria then a method produces different results) as well, see Hämäläinen and Alaja (2008).

An alternative approach for examination of the path and scale dependency of the BWM can be based on numerical simulations, which might be an interesting topic for the future research, see e.g. Lahtinen et al. (2020).

6 Conclusions

The aim of this paper was to examine the path and scale dependence of the (linearized) Best–Worst Method (BWM) and its accuracy in pairwise comparisons. For this purpose, an experiment with over 800 respondents was carried out. The respondents' task consisted of pairwise comparisons of the area (size) of six geometric figures, where the order of pairwise comparisons and the scale for comparisons differed across questionnaire forms.

The results of the experiment showed that the BWM is both path and scale dependent at the 0.001 significance level. Therefore, apart from the BWM's obvious advantages, a decision maker should be aware that the method, similarly to many other pairwise comparisons methods, also has its limitations (drawbacks).

Additionally, it was found that the most accurate estimations, on average, were obtained via continuous scale $[1, \infty)$, while the answers of respondents who used Saaty's integer and linguistic 9-point scales were slightly less precise. In addition, the mean relative error of estimations was below 18% for all geometric figures and all questionnaire forms, which can be considered a very favorable outcome expressing the strength of the pairwise comparisons method.

The design of the experiment didn't allow to determine the cause of the path and scale dependency of the BWM, hence it can be a subject of a further research. Also, future research can focus on the path and scale dependency of other pairwise comparisons methods, such as the analytic hierarchy process (AHP), AHP-Express, or the base criterion method (BCM).

Acknowledgements The research conducted by J. Mazurek was partially supported by the project GACR No. 21-03085S, Czech Republic. The research conducted by D. Strzałka was partially supported by the grant: Podpora mezinárodných mobilit na Slezské univerzite v Opave, No. CZ.02.2.69/0.0/0.0/18_053/0017871.

Data availability The datasets generated and analyzed during the current study are not publicly available due the fact that they constitute an excerpt of research in progress but are available from the corresponding author on reasonable request.

Declarations

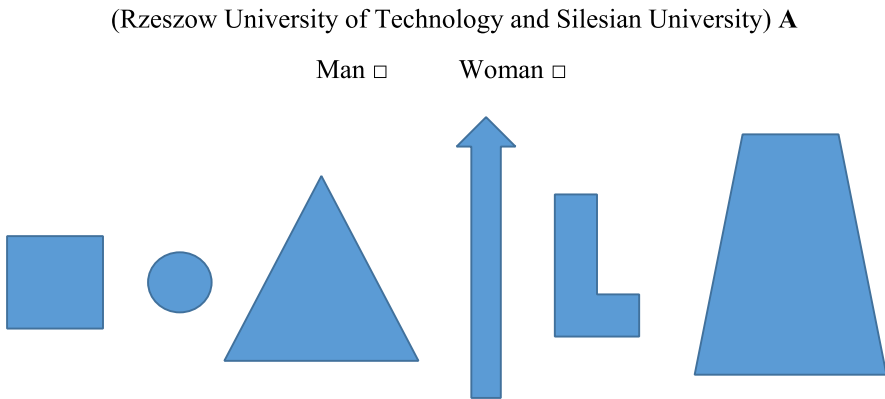
Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A: The questionnaire, form A

Path dependency project in pairwise comparisons

See Fig. 4.



Your task is to compare the size of the above geometric Figures. Of course, the trapezoid has the greatest area and the smallest one is the circle.

Compare the shapes areas with respect to the trapezoid (write how many times the area of the trapezoid is larger than the area of the other shapes, you can use integers and decimals from the range $<1, +\infty$), thus ≥ 1 . 1 denotes that in your opinion shapes areas are equal):

1. The area of a trapezoid is times greater than the area of a square.
2. The area of a trapezoid is times greater than the area of a circle.
3. The area of a trapezoid is times greater than the area of a triangle.
4. The area of a trapezoid is times greater than the area of an arrow.
5. The area of a trapezoid is times greater than the area of a L shape.

Now compare the sizes of the shapes with the circle in the same way (write how many times the area of the shape is larger than the area of the circle, you can use integers and decimals according to the rules above):

1. The area of square is grater than the area of a circle.
2. The area of triangle is grater than the area of a circle.
3. The area of arrow is grater than the area of a circle.
4. The area of L shape is grater than the area of a circle.

Fig. 4 Form A

References

- Abadia FA, Sahebib IG, Arabc A, Alavid A, Karachi H (2018) Application of best-worst method in evaluation of medical tourism development strategy. *Decis Sci Lett* 7:77–86
- Ahmadi AB, Kusi-Sarpong S, Rezaei J (2017) Assessing the social sustainability of supply chains using best worst method. *Resour Conserv Recycl* 126:99–106
- Allen P, Bennett K, Heritage B (2018) *SPSS statistics: a practical guide*, 4th edn. Cengage Learning, Melbourne
- Anderson DR, Sweeney DJ, Williams TA (1996) *Statistics for business and economics*, 6th edn. West Publishing Co., Minneapolis/St. Paul
- Barker HR, Barker BM (1984) *Multivariate analysis of variance (MANOVA): a practical guide to its use in scientific decision-making*. University of Alabama Press, Tuscaloosa
- Beemsterboer DJC, Hendrix EMT, Claassen GDH (2018) On solving the best-worst method in multi-criteria decision-making. *IFACPapersOnLine* 51(11):1660–1665. <https://doi.org/10.1016/j.ifacol.2018.08.218>
- Brocklesby J (2016) The what, the why and the how of behavioural operational research: an invitation to potential sceptics. *Eur J Oper Res* 249:796–805
- Brown CE (1998) *Applied multivariate statistics in geohydrology and related sciences*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-80328-4_7
- Brunelli M, Rezaei J (2019) A multiplicative best-worst method for multi-criteria decision making. *Oper Res Lett* 47(1):12–15
- Chang M-H, Liou JH, Lo H-W (2019) A hybrid MCDM model for evaluating strategic alliance partners in the green biopharmaceutical industry. *Sustainability* 11(15):4065. <https://doi.org/10.3390/su11154065>
- Finch H (2005) Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodol Eur J Res Methods Behav Soc Sci* 1(1):27–38. <https://doi.org/10.1027/1614-1881.1.1.27>
- Franco LA, Hämäläinen RP, Rouwette EA, Leppanen I (2021) Taking stock of behavioural OR: a review of behavioural studies with an intervention focus. *Eur J Oper Res* 293(2):401–418
- Franek J, Kresta A (2014) Judgment scales and consistency measure in AHP. *Procedia Econ Finance* 12:164–173. [https://doi.org/10.1016/S2212-5671\(14\)00332-3](https://doi.org/10.1016/S2212-5671(14)00332-3)
- Gupta H, Barua MK (2017) Supplier selection among SMEs on the basis of their green innovation ability using BWM and fuzzy TOPSIS. *J Clean Prod* 152:242–258. <https://doi.org/10.1016/j.jclepro.2017.03.125>
- Hämäläinen RP, Alaja S (2008) The threat of weighting biases in environmental decision analysis. *Ecol Econ* 68(1–2):556–569
- Hämäläinen RP, Lahtinen TJ (2016) Path dependence in operational research—how the modeling process can influence the results. *Oper Res Perspect* 3:14–20. <https://doi.org/10.1016/j.orp.2016.03.001>
- Hämäläinen RP, Luoma J, Saarinen E (2013) On the importance of behavioral operational research: the case of understanding and communicating about dynamic systems. *Eur J Oper Res* 228(3):623–634
- Harker PT, Vargas LG (1987) The theory of ratio scale estimation: Saaty’s analytic hierarchy process. *Manag Sci* 33(11):1383–1403
- Ishizaka A, Balkenborg D, Kaplan T (2010) Influence of aggregation and measurement scale on ranking a compromise alternative in AHP. *J Oper Res Soc* 62(4):700–710
- Knief U, Forstmeier W (2021) Violating the normality assumption may be the lesser of two evils. *Behav Res* 53:2576–2590. <https://doi.org/10.3758/s13428-021-01587-5>
- Koczkodaj WW, Liu F, Marek VW, Mazurek J, Mazurek M, Mikhailov L, Zel C, Pedrycz W, Przelaskowski A, Schumann A, Smarzewski R, Strzalka D, Szybowski J, Yayli Y (2020) On the use of group theory to generalize elements of pairwise comparisons matrix: a cautionary note. *Int J Approx Reason* 124:59–65. <https://doi.org/10.1016/j.ijar.2020.05.008>
- Krider RE, Raghurir P, Krishna A (2001) Pizzas: π or square? Psychophysical biases in area comparisons. *Market Sci* 20(4):405–425. <https://doi.org/10.1287/mksc.20.4.405.9756>
- Kunc M, Malpass J, White L (2016) *Behavioral operational research: theory, methodology and practice*. Palgrave Macmillan, London
- Lahtinen TJ, Hämäläinen RP (2016) Path dependence and biases in the even swaps decision analysis method. *Eur J Oper Res* 249(3):890–898. <https://doi.org/10.1016/j.ejor.2015.09.056>
- Lahtinen TJ, Guillaume JH, Hämäläinen RP (2017) Why pay attention to paths in the practice of environmental modelling? *Environ Model Softw* 92:74–81

- Lahtinen TJ, Hämäläinen RP, Jenytn C (2020) On preference elicitation processes which mitigate the accumulation of biases in multicriteria decision analysis. *Eur J Oper Res* 282(1):201–210
- Landry M, Malouin J-L, Oral M (1983) Model validation in operations research. *Eur J Oper Res* 14(3):207–220. [https://doi.org/10.1016/0377-2217\(83\)90257-6](https://doi.org/10.1016/0377-2217(83)90257-6)
- Latimer C, Stevens C, Irish M, Webber L (2000) Attentional biases in geometric form perception. *Q J Exp Psychol Sect A* 53(3):765–791. <https://doi.org/10.1080/713755915>
- Leskinen P (2008) Numerical scaling of ratio scale utilities in multicriteria decision analysis with geometric model. *J Oper Res Soc* 59(3):407–415
- Ma D, Zheng X (1991) 9/9-9/1 scale method of AHP. In: 2nd international symposium on AHP, vol 1. University of Pittsburgh, Pittsburgh, p 197202
- Mazurek J (2023) Advances in pairwise comparisons: detection, evaluation and reduction of inconsistency, multiple criteria decision making series. Springer Nature, Switzerland
- Mazurek J, Perzina R, Ramik J, Bartl D (2021) A numerical comparison of the sensitivity of the geometric mean method, eigenvalue method, and best worst method. *Mathematics* 9:554. <https://doi.org/10.3390/math9050554>
- Mi X, Tang M, Liao H, Shen W, Lev B (2019) The state-of-the art survey on integrations and applications of the best worst method in decision making: Why, what, what for and what's next? *Omega* 87:205–225. <https://doi.org/10.1016/j.omega.2019.01.009>
- Olson CL (1976) On choosing a test statistic in multivariate analysis of variance. *Psychol Bull* 83:579–586. <https://doi.org/10.1037/00332909.83.4.579>
- Poyhonen M, Hämäläinen RP (2001) On the convergence of multiattribute weighting methods. *Eur J Oper Res* 129(3):569–585
- Rezaei J (2015) Best-worst multi-criteria decision-making method. *Omega* 53:40–57
- Rezaei J (2016) Best-worst multi-criteria decision-making method: some properties and a linear model. *Omega* 64:126–130
- Rezaei J, Nispeling T, Sarkis J, Tavasszy L (2016) A supplier selection life cycle approach integrating traditional and environmental criteria using the best worst method. *J Clean Prod* 135:577–588. <https://doi.org/10.1016/j.jclepro.2016.06.125>
- Saaty TL (1977) A scaling method for priorities in hierarchical structures. *J Math Psychol* 15:234–281
- Saaty TL (1980) Analytic hierarchy process. McGraw-Hill, New York
- Salo A, Hämäläinen R (1997) On the measurement of preference in the analytic hierarchy process. *J Multi-Criteria Decis Anal* 6(6):309–319
- Setyono RP, Sarno R (2018) Vendor track record selection using best worst method. In: 2018 international seminar on application for technology of information and communication, Semarang, Indonesia, IEEE, pp 41–48. <https://doi.org/10.1109/ISEMANTIC.2018.8549711>
- Thurstone LL (1927) A law of comparative judgments. *Psychol Rev* 34:273–286
- Warne RT (2014) A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Pract Assess Res Eval* 19(17):1–10
- Weinfurt KP (1995) Multivariate analysis of variance. In: Grimm LG, Yarnold PR (eds) Reading and understanding multivariate statistics. American Psychological Association, Washington, D.C., pp 245–276
- Zaointz C (2022) Real statistics using excel. Available from: <https://www.real-statistics.com/multivariate-statistics/multivariateanalysis-of-variance-manova/>
- Zientek LR, Thompson B (2009) Matrix summaries improve research reports: secondary analyses using published literature. *Educ Res* 38(5):343–352. <https://doi.org/10.3102/0013189X09339056>