

An experimental approach: Converting verbal expressions to numerical scales

Zsombor Szádóczki ^{a,b}, Sándor Bozóki ^{a,b}, László Sipos ^{c,d}, Zsófia Galambosi ^c

^a Research Group of Operations Research and Decision Systems, Research Laboratory on Engineering & Management Intelligence, HUN-REN Institute for Computer Science and Control (SZTAKI), 1111 Kende u. 13-17., Budapest, Hungary

^b Department of Operations Research and Actuarial Sciences, Corvinus University of Budapest, 1093 Fővám tér 8., Budapest, Hungary

^c Hungarian University of Agriculture and Life Sciences, Institute of Food Science and Technology, Department of Postharvest, Commercial and Sensory Science, 1118, Villányi út 29-43., Budapest, Hungary

^d Institute of Economics, ELTE Centre for Economic and Regional Studies (ELTE CERS), Eötvös Loránd University, 1097 Tóth Kálmán utca 4., Budapest, Hungary

ARTICLE INFO

Keywords:

Multicriteria decision making
Decision analysis
Preference modeling
Pairwise comparisons
AHP
Scales

ABSTRACT

A key issue in decision problems is the selection and use of the appropriate response scale. In this paper verbal expressions are converted into numerical scales for a subjective problem instance. In our experiment, we conducted a color selection test with 462 subjects by testing six colors on color-calibrated tablets in ISO standardized sensory test booths. The colors were evaluated both in a pairwise comparison matrix (indirect ranking with four-item verbal category scale) and on a direct scoring basis. We determined scales that provide the closest results on average and individually to the direct scoring, based on the eigenvector and the logarithmic least squares methods. All results show that the difference between verbal expressions is much smaller than the one used by most of the common numerical scales. As a sensitivity analysis, different discrepancy measures were investigated, and further experiments were also carried out. All the main findings were confirmed to be robust regarding the weight calculation technique, the discrepancy measure, and the problem instance as well. The results suggest that several methods (e.g., the AHP, and different ISO standards) need substantial revision. The respondents' inconsistency was also analyzed with a repeated question regarding their preference between a given pair of colors. It is shown that most decision makers answer similarly for the second time, but there can be significant (even ordinal) differences. The respondents whose answers are further from the original tend to be more inconsistent in general.

1. Introduction

Pairwise comparisons are popular in decision making problems [1–3], preference modeling [4], supply chain management [5], as well as in sports [6]. One of the most frequently used multicriteria decision making (MCDM) methods is the Analytic Hierarchy Process (AHP) proposed by Saaty [7,8], which applies pairwise comparison matrices (PCMs). The appropriate entry of a (multiplicative) PCM shows the decision maker's relative preference of the alternative in the row over the alternative in the column.

Empirical pairwise comparison matrices are collected from real or experimental decision problems [9,10]. They can substantially differ from simulated or specifically chosen example matrices that are mainly considered in the existing literature [11]. Thus, their analysis can lead to important consequences and practical recommendations in decision support.

In the classical AHP, the so-called Saaty-scale or fundamental scale—i.e., the discrete scale of $\{1/9, 1/8, \dots, 1/2, 1, 2, \dots, 8, 9\}$ —is applied to make each comparison between the pairs of alternatives. However, these values are converted from the verbal assessments of the decision makers chosen from an appropriate discrete list of possibilities.

The data provided by empirical pairwise comparison matrices and alternative methods of preference elicitation, such as direct scoring, can be used to verify in what extent does a given scale correspond to the decision makers' judgments.

In this paper, we conduct sensory-based experiments with pairwise comparisons to address the discrepancy between the commonly used scales and reality. The preferences of university students regarding different colors are collected via pairwise comparisons based on a four-item verbal category scale as well as direct scoring. The empirical distributions of the comparisons calculated from direct evaluation

* Correspondence to: 1111 Kende u. 13-17., Budapest, Hungary.

E-mail addresses: szadoczki.zsombor@sztaki.hu (Z. Szádóczki), bozoki.sandor@sztaki.hu (S. Bozóki), sipos.laszlo@uni-mate.hu (L. Sipos), galambosi.zsofia@uni-mate.hu (Z. Galambosi).

<https://doi.org/10.1016/j.orp.2026.100416>

Received 13 March 2026; Received in revised form 8 July 2026; Accepted 21 August 2026

Available online 1 September 2026

2214-7160/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

scores, and the verbal category scale applied for pairwise comparisons are analyzed to determine the optimal numerical scale that should be used between the conversion of verbal and numerical scales of pairwise comparisons in the given problem. Both the individually optimal scales, and the one that is the best on average are determined using the Eigenvector and the Logarithmic Least Squares Methods based on the Euclidean distances calculated with the direct evaluation scores. One of the questions regarding the preference between a pair of colors was repeated at the end of the experiment that also led to interesting findings regarding the inconsistency of the individuals.

It is important to mention that direct scoring is not considered to be the ground truth, it is a simple benchmark method that is necessary to evaluate a subjective problem instance. Naturally, several other benchmark methods could have been applied. Direct scoring was chosen mainly because its simplicity and extensive practical usage. However, it has its own drawbacks, and it is also subject to noise, instability, and scale effects [12,13].

As a robustness analysis, first, different discrepancy measures were applied instead of the Euclidean distance. Then, several further experiments were carried out with different samples of individuals and different entities to compare as well. These experiments corroborate all the findings presented in connection with the color-choice test. The results suggest that several multicriteria decision making methods (e.g., the AHP), and different ISO standards [14] need substantial revision.

The rest of the paper is organized as follows. Section 2 contains a literature review related to converting verbal scales to numerical ones in decision making problems. The methodology of the questionnaires and their evaluation are detailed in Section 3. Section 4 presents the main results, the relation of direct scoring and pairwise comparisons, as well as the analysis of the respondents' inconsistency. Section 5 shows the results of a comprehensive robustness check. First, Section 5.1 presents the analysis using different discrepancy measures, while Section 5.2 describes further experiments. Finally, Section 6 concludes and raises further research questions.

2. Literature review

Empirical pairwise comparison matrices can significantly differ from simulated ones, and form an essential source of research. Bozóki et al. [9] examine 454 PCMs (with 227 decision makers) and analyze the effect of the size, the question ordering, and the type of the problem (subjective or objective) on the inconsistency of the matrices. Cavallo and Ishizaka [13] compare the precision of more than 100 respondents when they evaluate the area of nine geometric figures and five distances between cities applying pairwise comparisons, direct scoring, and the best–worst method [15].

Huizingh and Vrolijk [16] applies a laboratory study with 180 participants to compare the results of the Analytic Hierarchy Process using numerical values and verbal expressions converted by the fundamental scale $(\{1, 2, \dots, 9\})$. The results show that the fundamental scale tends to overestimate the differences in preference, but the numerical mode only shows slightly better results (regarding inconsistency and the finally chosen alternative).

Pöyhönen et al. [17] perform a comparative study applying an experiment on areas of figures with different shapes. They compare three different scales (linear (fundamental), $9/9 - 9/1$, and balanced). The main finding of the experiment is that the perceived meaning of the verbal expressions varies from one decision maker to the other, and also depends on the set of elements involved in the comparison. They show that alternative numerical scales yield more accurate estimates than the usual 1-to-9 scale, and these also reduce the inconsistency of the comparison matrices measured by the consistency measure (CM) that is independent from the choice of the scale [18,19].

Elliott [20] examines three different scale types, and based on the answers of 64 respondents, finds that none of these scales can accurately capture the preferences of all individuals. On a real engineering

decision problem it is shown that the choice of scale can also affect the preferred decision. Finan and Hurley [21] apply experiments to argue for the calibration of the applied scales, particularly proposing a geometric scale based on a single parameter.

Dong et al. [22] propose to generate the numerical scale of each individual based on a linguistic model in order to handle the problem caused by the fact that the individuals understand the same verbal scale in a different manner. They apply nonlinear programming to determine the numerical values based on five different scales known in the literature.

Rokou and Kirytopoulos [23] show that the problem of choosing the appropriate scale is essential in group decision making. They present a group calibration process that adjusts individuals' preferences based on their answers on a set of standardized questions. Meesariganda and Ishizaka [24] compare eight different numerical scales. First, they apply the verbal scales to compare alternatives with known measures, e.g., the surface of figures. Then, based on the Euclidean distances, the best matching scale representing the real values is selected for each individual. Ahmed and Kilic [25] propose a scale individualization approach based on compatibility. Dutta et al. [26] recommend an approach comprising two steps to identify the most suitable personalized numerical scale for individuals.

Koczkodaj et al. [27] propose a method to reduce the number of rating scale items without the loss of predictability, while Koczkodaj et al. [28] apply differential evolution optimization to improve rating scale predictability.

Franek and Kresta [29] compare eight different scales regarding the resulting priorities and consistency on an example problem. Results suggest that judgment scales have a significant impact on criteria priorities but not on the ranking of criteria. The inconsistency varies among applied judgment scales for which the appropriate random indices have been calculated by the authors (see Franek and Kresta [29, Table 7]) to determine the suitable CR inconsistency values. Goepel [30] proposes three different scales (generalized balanced, adaptive, and adaptive-balanced scales) also depending on the number of examined criteria, and provides a thorough comparative analysis as well.

Cavallo and Ishizaka [10] conduct experiments, where the decision makers asked to compare the distance of different cities. Three problems are considered with 52 PCMs each, where eight different scales examined earlier in the literature are compared to each other. They find that the inverse linear scale provides the best results on average in every problem.

Table 1 contains some of the most popular scales proposed in the literature to assess verbal statements for pairwise comparison matrices. It is worth mentioning that even though all of these scales are presented in a general framework, they are used with a given parameter combination (shown in the 'Usual choice of parameters' column in Table 1) in the vast majority of articles, and in practice as well.

Most of the studies comparing different scales focus on objective problems, e.g., evaluating the size of figures. They also deal with the usual parameter combinations of the different scales. This paper examines a large number of scales using a subjective problem, i.e., comparing different colors, which could be significantly different from objective tasks.

3. Methodology

3.1. Pairwise comparison matrices

In this study, the concept of pairwise comparison matrices and related tools are widely used. Let us denote the number of alternatives by n .

Definition 1 (*Pairwise Comparison Matrix (PCM)*). The $n \times n$ matrix $A = [a_{ij}]$ is a pairwise comparison matrix, if it is

Table 1
Popular scales proposed in connection with pairwise comparison matrices.

Name	Description	Parameters	Usual choice of parameters	Scale values	Reference
Linear	$\alpha \cdot x$	$\alpha > 0, x \in \{1, 2, \dots, 9\}$	$\alpha = 1$	$\{1, 2, \dots, 9\}$	[7]
Affine	$\alpha \cdot x + \beta$	$\alpha > 0, \beta > 0, x \in \{1, 2, \dots, 9\}$	$\alpha = 0.1, \beta = 1$	$\{1.1, 1.2, \dots, 1.9\}$	[31]
Power	x^α	$\alpha > 1, x \in \{1, 2, \dots, 9\}$	$\alpha = 2$	$\{1, 4, 9, \dots, 81\}$	[32]
Root	$\sqrt[x]{x}$	$\alpha > 1, x \in \{1, 2, \dots, 9\}$	$\alpha = 2$	$\{1, \sqrt{2}, \sqrt{3}, \dots, 3\}$	[32]
Geometric	α^{x-1}	$\alpha > 1, x \in \{1, 2, \dots, 9\}, x \in \{1, 1.5, \dots, 4\}, \text{ etc.}$	$\alpha = \sqrt{2}$ or $\alpha = 2$	$\{1, \sqrt{2}, 2, \dots, 16\}$ or $\{1, 2, 4, \dots, 256\}$	[33]
Inverse linear	$\frac{9}{10-x}$	$x \in \{1, 2, \dots, 9\}$	–	$\{1, 1.13, 1.29, \dots, 9\}$	[34]
Asymptotic	$e^{\tanh^{-1}\left(\frac{\sqrt{3}(x-1)}{14}\right)}$	$x \in \{1, 2, \dots, 9\}$	–	$\{1, 1.13, 1.29, \dots, 13.93\}$	[35]
Balanced	$\frac{x}{1-x} \left(= \frac{9+y}{11-y} \right)$	$x \in \{0.5, 0.55, \dots, 0.9\}, y \in \{1, 2, \dots, 9\}$	–	$\{1, 1.22, 1.5, \dots, 9\}$	[36]
Balanced power	$9^{\frac{x-1}{n-1}}$	$x \in \{1, 2, \dots, n\}$	$n = 9$	$\{1, 1.32, 1.73, \dots, 9\}$	[20]
Logarithmic	$\log_\alpha(x + \alpha - 1)$	$\alpha > 1, x \in \{1, 2, \dots, n\}$	$\alpha = 2$	$\{1, 1.58, 2, \dots, 3.32\}$	[37]
Koczkodaj	$1 + \frac{x-1}{n-1}$	$x \in \{1, 2, \dots, n\}$	$n = 9$	$\{1, 1.125, 1.25, \dots, 2\}$	[38]

- positive ($a_{ij} > 0 \forall i, j$) and
- reciprocal ($1/a_{ij} = a_{ji} \forall i, j$).

One of the most commonly analyzed questions is the (in)consistency of PCMs.

Definition 2 (Consistent PCM). A PCM is consistent if $a_{ik} = a_{ij}a_{jk} \forall i, j, k$. If a PCM is not consistent, then it is called inconsistent.

The level of inconsistency can be measured several ways [39], and there are different systems to validate such inconsistency metrics [40]. However, in practice, still the Consistency Ratio [7] is the most commonly used index for these purposes. It is also related to the 10% rule (or 0.1 rule), which states that a decision maker should be asked to reconsider a PCM in the case when the corresponding CR is larger than 10%.

Definition 3 (Consistency Ratio (CR)). The consistency ratio (CR) of an $n \times n$ PCM A is determined by Eq. (1).

$$CR = \frac{CI}{RI}, \quad (1)$$

where CI denotes the Consistency Index, which is defined by Eq. (2).

$$CI = \frac{\lambda_{\max} - n}{n - 1}, \quad (2)$$

where $\lambda_{\max}$ is the principal eigenvalue of matrix A, and RI is the Random Index, which is the average CI obtained from a sufficiently large set of randomly generated PCMs of size n using the fundamental scale ($\{1, 2, \dots, 9\}$) and discrete uniform distribution.

It is common to calculate a weight vector from a given PCM that represents the estimation of the underlying preferences of the decision maker. Two of the most commonly used weight calculation techniques are the Eigenvector Method [7], and the Logarithmic Least Squares (Geometric Mean) Method [41].

Definition 4 (Eigenvector Method (EM)). The weight vector $\mathbf{w}$ of an $n \times n$ PCM A determined by the EM is defined by Eq. (3).

$$A\mathbf{w} = \lambda_{\max}\mathbf{w}, \quad (3)$$

where the componentwise positive principal eigenvector $\mathbf{w}$ is unique up to a scalar multiplication.

Definition 5 (Logarithmic Least Squares Method (LLSM)). The weight vector $\mathbf{w}$ of an $n \times n$ PCM A determined by the LLSM is given by Eq. (4).

$$\min_{\mathbf{w}} \sum_{i=1}^n \sum_{j=1}^n \left(\ln(a_{ij}) - \ln\left(\frac{w_i}{w_j}\right) \right)^2, \quad (4)$$

where w_i is the i th coordinate of $\mathbf{w}$.

3.2. Framework of the ratio scale generation and comparison

In an experiment, it is interesting to examine the scale that provides the best results on average. Where ‘the best’ in our methodology means the closest (based on the Euclidean distance) to the objective values for an objective analysis, and the closest to the values calculated from other (benchmark) preference elicitation methods, such as the direct scoring, in the case of subjective problems. However, it also makes sense to find the optimal scale for each individual as well.

In order to answer these questions in our experiment, the EM and LLSM weight vectors of the individual pairwise comparison matrices have been calculated for several scales, applying the following parametrization of the used four-item verbal scale.

- 1 – equally like;
- S – like it a little bit more;
- M – like it moderately more;
- L – like it much more.

Where $1 < S < M < L$, thus a strict relation is assumed between the different levels, and the category representing indifference (‘equally like’) must be converted to 1 as a numerical value.

All parameter combinations were tested on a grid with a step size of 0.1 in each variable, where the above assumptions hold, and $S \leq 5$, $M \leq 10$, and $L \leq 15$. That is a total of 236 880 parameter combinations.

3.3. Conducted experiments

In the conducted color-choice test, the colors selected were those that were clearly distinguishable on the subjects’ mental maps: red, green, blue, magenta, turquoise and yellow. The RGB coordinates of these color names can be found in the Color Namer database (<https://colornaming.net/#colour-namer>): red (R:189, G:62, B:57), green (R:90, G:151, B:90), blue (R:84, G:110, B: 183), magenta (R: 179, G: 55, B: 151), turquoise (R: 63, G: 185, B: 177), and yellow (R: 227, G: 203, B: 78) [42]. The tests were conducted on color-calibrated tablets (Samsung Galaxy Tab A 2018, 10.5) with identical test geometry and maximum brightness settings in the sensory test booths of the Sensory Laboratory, Institute of Food Science and Technology, Hungarian University of Agricultural and Life Sciences. The laboratory conforms to the International Standard for the design requirements for sensory test [43].

In the test, each color was initially compared with every other color in pairs based on their stated preferences, resulting in a total of 15 pairs

(calculated as $n(n-1)/2$). In the first phase, participants were first asked to indicate their preferred color in pairs based on their preference, and then to rate their level of preference using a four-item verbal category scale: equally like, like it a little bit more, like it moderately more, like it much more (see Fig. A.1 in Appendix). In the second phase of the experiment, the participants were asked to rate the colors on an eleven-point scale, with 0 indicating a strong dislike and 10 indicating a strong preference (see Figs. A.2 and A.3 in Appendix). Consequently, the colors were evaluated both on a pairwise comparison (indirect ranking) and on an individual basis (direct ranking). In the third phase, participants were asked a series of socio-demographic questions. These included queries regarding gender, age, and county. To assess the consistency of the responses, the question regarding the preference of the secondly asked pair of colors was repeated at the conclusion of the test in reverse order.

4. Results

There were a total of 462 respondents in our color-choice test. However, before the calculations, it was necessary to clean the dataset as follows.

1. The results of the individuals who did not use the whole verbal scale (e.g., there were no pair of colors for which they liked one a little bit more than the other) were removed. This could have caused problems when the individually best scales were determined as there would have been a number of optimal cases. For instance, in the previous example S is not used, thus, M and L are optimized, and S can take any value that satisfies the assumptions. This results in the removal of the answers of 98 respondents.
2. It is easy to compare the numerical scale values with the ratio of the direct scoring results. However, it was possible to determine 0 as the score of a color. The answers of the 63 respondents who used this value were also removed from the calculations.

After that a total of 301 PCMs and direct scoring results were evaluated. It is also worth mentioning that we did calculate the results without excluding these respondents (using transformations for the second point), and all the main findings were the same.

4.1. Optimal scales

Fig. 1 shows the histograms of the 1, S , M , and L parameters from the direct evaluations, respectively, i.e., the histograms of the ratios of directly evaluated values, where the pairwise comparison resulted as ‘equally like’ (indifference), ‘like it a little bit more’, ‘like it moderately more’, and ‘like it much more’ for the given two colors, respectively.

One can see that in the first histogram which is representing the indifference between the two colors, 1 is the most common value of the ratio of direct evaluation scores. The distribution is more or less symmetric to that in the sense that the ratio value is smaller than 1 approximately the same number of times as it is greater than 1.

Then, as we go on to the more extreme preference values (like it a little bit more, moderately more, and much more) the common values for the ratio of direct evaluation scores starting to be larger and larger as one would expect. However, interestingly enough, there are still smaller than 1 ratios for the most extreme case as well, showing the inconsistency of the respondents.

Here we only present the results for those parameter combinations, where $S, M, L \leq 4$, while all the other assumptions hold, as it turned out that the scales including larger values are almost never optimal for any individuals. Figs. 2 and 3 show the number of individuals for whom a given scale has been optimal based on the Euclidean distance using the Eigenvector and the Logarithmic Least Squares weight calculation techniques, respectively. The greater number of individuals is highlighted by the color and the size of the given point as well.

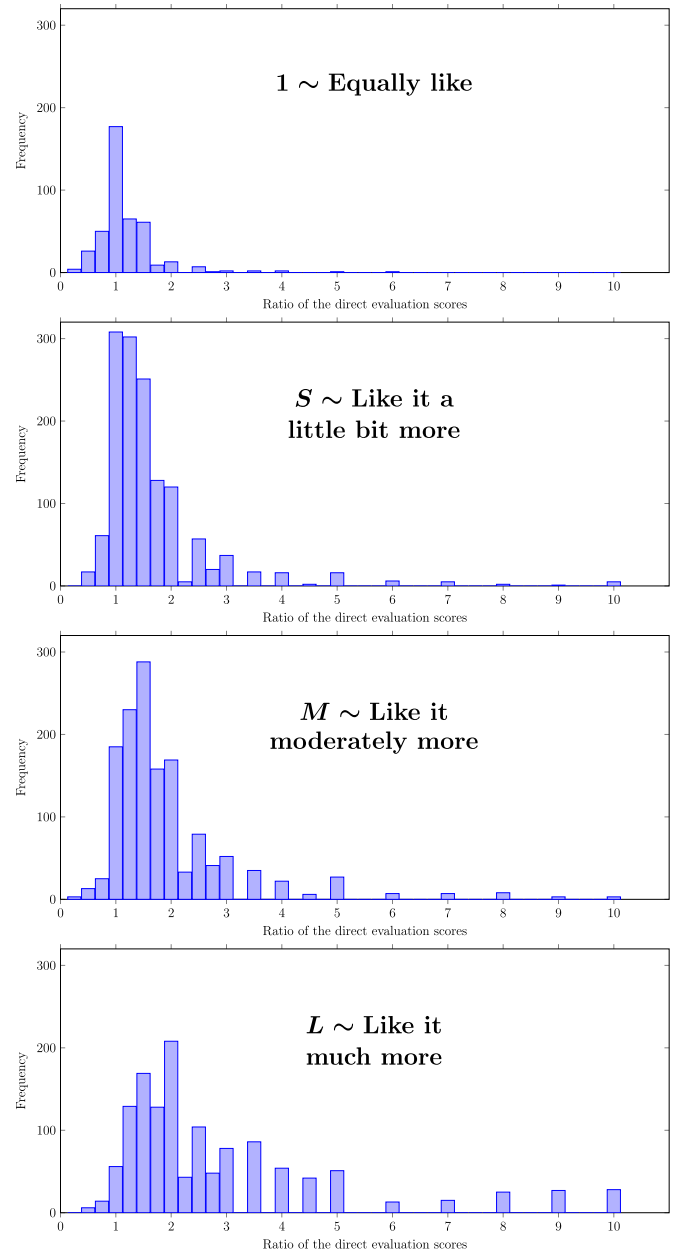


Fig. 1. The histograms of the 1, S , M , and L values from the direct evaluation, respectively, i.e., the histograms of the ratios of directly evaluated values, where the pairwise comparison resulted as ‘equally like’ (indifference), ‘like it a little bit more’, ‘like it moderately more’, and ‘like it much more’ for the given two colors, respectively.

As one can see, the results of the two weight calculation methods are similar, and the most important conclusions are the same for both. All the commonly optimal scales include smaller than 3 parameter values for even L . The scale that results in the smallest average Euclidean distance uses the parameter combination of $S = 1.5$, $M = 1.7$, and $L = 2$ for both weight calculation techniques.

However, it is important to highlight the fact that the parameter combinations that are relatively close to the one that is proved to be optimal on average according to the Euclidean distance, all provide close-to-optimal values. Meanwhile, the average distance for the scales that include larger values (≥ 3) are far from being optimal.

The optimal scales of individuals for the EM

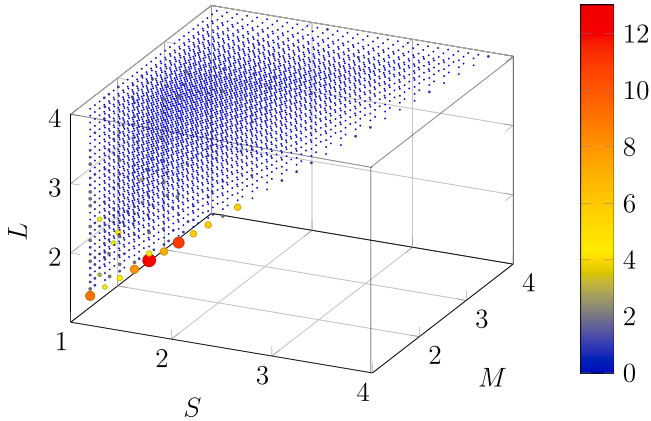


Fig. 2. The optimal scales of individuals based on the Eigenvector Method, $1.1 \leq S \leq 3.8$, $1.2 \leq M \leq 3.9$, $1.3 \leq L \leq 4$, and $S < M < L$.

The optimal scales of individuals for the LLSM

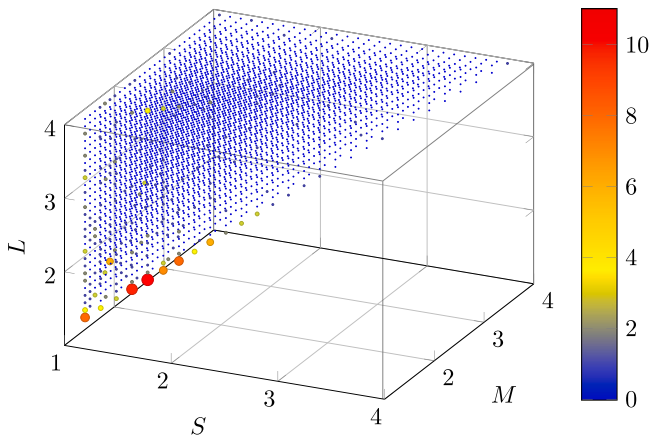


Fig. 3. The optimal scales of individuals based on the Logarithmic Least Squares Method, $1.1 \leq S \leq 3.8$, $1.2 \leq M \leq 3.9$, $1.3 \leq L \leq 4$, and $S < M < L$.

4.2. Comparison with existing scales

The obtained optimal scale is compared with the previously presented existing scales (Table 1) based on the Euclidean distances between the normalized weight vectors (whose elements sum up to 1) calculated from pairwise comparisons and direct scoring in Table 2.

It is important to highlight that in the experiments of the current study, a four-item scale is used, while the already existing scales apply nine possible values. However, it has been shown in the literature that odd numbers are significantly more often used in empirical pairwise comparison matrices than even numbers. Gass and Standard [44] reported that odd numbers are twice as often chosen based on 384 matrices, while the contrast found to be even stronger by Poesz [45] based on 137 PCMs, i.e., odd numbers were used roughly four times as often as even ones. A reason behind this could be the fact that the AHP puts explicit word descriptions only on the odd numbers. It is also shown by both above-mentioned studies that extreme comparison values (e.g., 8 and 9) are especially rarely used (approximately 7 percent of the comparisons resulted in an 8 or 9 value combined in both investigations).

Based on these results, the already existing scales were modified to make it possible to have a fair comparison with the scales examined by the current study. Namely, in the case of the most commonly used Saaty

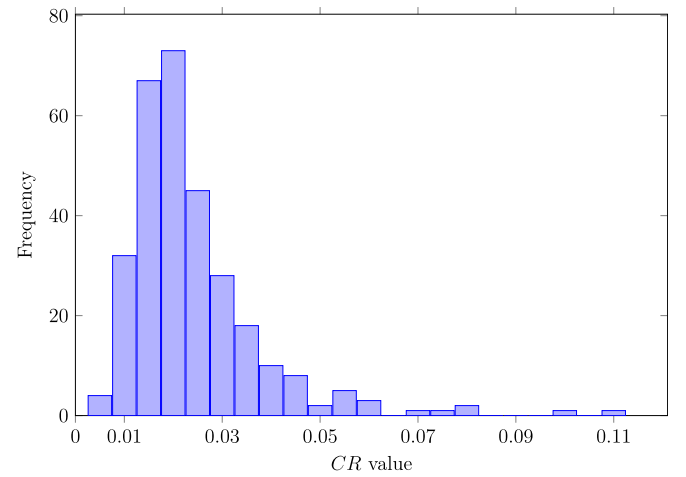


Fig. 4. The histogram of the CR values based on the $S = 1.5$, $M = 1.7$, and $L = 2$ scale.

(linear) scale, only the odd numbers were used, and the most extreme evaluation was omitted from the analysis. Thus, the resulting four-element scale is {1, 3, 5, 7}. A similar logic was applied to all existing scales, only the first, third, fifth, and seventh element was used for the comparison presented in Table 2.

As one can see the scales applying relatively high values (e.g., the linear, power, and geometric scales) tend to perform rather poorly. While the ones with smaller, denser items (for instance, the affine, root, inverse linear, asymptotic, and Koczkodaj scales) provide reasonable results. This is in line with the aforementioned fact that the scales that are relatively close to the one that is proved to be optimal on average according to the Euclidean distance tend to provide close-to-optimal values. In a sense, based on the results, the existing scales benefit from the omission of the most extreme values. Nevertheless, the best performing alternative is the empirically optimal scale. As shown earlier, individually optimal scales tend to disperse around the scale that is optimal on average, thus, the individual analysis shows the exact same pattern.

It is also worth highlighting that the linear (Saaty) scale that is one of the most frequently applied option in practice, is among the worst performing scales regarding the average Euclidean distance. Its provided distance is approximately three times as large as the one belonging to the empirically optimal scale. The magnitude of this Euclidean distance (0.2245) is roughly the same as the Euclidean distance (0.2236) between weight vectors $w = (0.45, 0.3, 0.15, 0.1)$ and $w' = (0.3, 0.45, 0.1, 0.15)$. A robustness analysis for other metrics is conducted in Section 5.1 that shows the results are not heavily dependent on the chosen discrepancy measure (see also Table A.1).

4.3. The inconsistency of participants

One can see the histogram of the respondents' inconsistencies measured by the CR index in Fig. 4.

However, the CR consistency index heavily builds on the fundamental scale with the generation of the RI. For a 6×6 matrix, the original RI is 1.249 calculated by simulations in several studies (see, for instance, Bozókí and Rapcsák [46] or Ágoston and Csátó [47]). We ran a simulation with 10 million randomly generated PCMs using the scale which resulted in the smallest average Euclidean distance ($S = 1.5$, $M = 1.7$, and $L = 2$). The calculated modified random index (let us denote it by $\widehat{RI}$) is 0.09224. Thus, a modified $\widehat{CR}$ can be calculated using $\widehat{RI}$, where the original CR index is multiplied by ~ 13.5 .

Table 2

Comparison with existing scales using the Eigenvector Method for weight calculation. Each scale is obtained by using the usual choice of parameters presented in Table 1. For the geometric scale, the parameter choice that provides the better results, $\alpha = \sqrt{2}$, is applied. For the affine scale, always the alternative that is first in the (red, green, blue, magenta, turquoise, yellow) list was marked as preferred with 1.1 value, if the decision maker indicated indifference between the given pair.

Name	1	S	M	L	Average Euclidean distance
Empirically optimal scale	1	1.5	1.7	2	0.0800
Linear	1	3	5	7	0.2245
Affine	1.1	1.3	1.5	1.7	0.0855
Power	1	9	25	49	0.4426
Root	1	1.73	2.24	2.65	0.0941
Geometric	1	2	4	8	0.2191
Inverse linear (9/9 – 9/1)	1	1.29	1.8	3	0.0942
Asymptotic	1	1.29	1.72	2.6	0.0864
Balanced	1	1.5	2.33	4	0.1239
Balanced power	1	1.73	3	5.2	0.1612
Logarithmic	1	2	2.58	3	0.1101
Koczkodaj	1	1.25	1.5	1.75	0.0849

This shows the importance of the numerical scale in evaluating inconsistency as well. However, we do not recommend the application or generalization of the common 10% rule in this case. The problem is highly subjective, and of a larger size (6×6 PCM). Furthermore, the number of options is also much smaller than the ones applied in the fundamental scale resulting in differences of the measured inconsistency.

To assess the consistency of the respondents in a different way, the preference elicitation of the secondly asked pair was repeated at the conclusion of the test in reverse order.

Fig. 5 shows the CR inconsistencies for the previously examined PCMs via Box plots for the possible categories of individuals depending on how far their repeated answer was from the original, as well as the number of individuals in the different categories.

One can see that the measured inconsistency tends to be higher as the repeated answer is further from the original one. There are much less individuals who made totally different judgments for the second time compared to their original answer. More than 77% of the respondents (232/301) made the same judgment for the second time as well, or only modified their answer with one step in the verbal scale. However, we can even see one respondent, who reversed the most extreme preference between the pair. For that individual, only a point measuring the corresponding CR is displayed on Fig. 5 instead of a Box plot.

It is also worth mentioning that if there are at least 3 steps in the four-item verbal scale between the original and the repeated preference, that means even the ordinal preference (which color is preferred) is changed at least to indifference. If the number of steps is even larger than 3, then the preference must be reversed. However, the number of individuals whose decisions resulted in a difference of at least 3 steps are a little bit less than 10% of the respondents (30/301).

There are two interesting findings based on the repeated preference elicitation.

1. Even though most of the respondents ($\sim 77\%$) answered almost the same way, there can be significant (even ordinal) differences in the results depending on when we ask the decision maker a given question.
2. The individuals who are inconsistent regarding their (repeated) answer to a question tend to be more inconsistent in general.

5. Robustness analysis

A twofold robustness analysis of the results was performed in order to demonstrate the generality of the findings. First, the sensitivity of the results to the chosen discrepancy metric is analyzed. Then, the robustness regarding the problem instance is shown via further, sufficiently different, experiments.

5.1. Sensitivity to the applied discrepancy metric

All the previously presented results depend on the Euclidean distance between the weight vector obtained from pairwise comparisons and the vector derived from direct scoring. However, several other discrepancy measures are applied in preference analysis. Therefore, a sensitivity analysis was performed on the data regarding the metric of discrepancy, which included the following measures estimating cardinal difference.

- The most commonly used further distance metrics, such as, the Chebyshev (d_c), and Manhattan distances (d_m);
- similarity or compatibility indices (typically ratio-based discrepancy measures), such as, the cosine and dice similarity indices (C and D , respectively [48,49]), as well as the Jaccard (J) and Garuti (G) compatibility indices [50,51].

Eqs. (5)–(10) show the formulas of the above-mentioned metrics, respectively, where w denotes the weight vector obtained from pairwise comparisons and v is the weight vector calculated from direct scoring. w and v are normalized by $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n v_i = 1$, respectively, and v_i and w_i denote the i th element of the appropriate vectors. In the case of the distances, the lower level is preferred, while for the similarity and compatibility measures (all of which has a range of $[0, 1]$), the higher value is better.

$$d_c(w, v) = \max_{i \in \{1, \dots, n\}} |w_i - v_i| \quad (5)$$

$$d_m(w, v) = \sum_{i=1}^n |w_i - v_i| \quad (6)$$

$$C(w, v) = \frac{\sum_{i=1}^n w_i v_i}{\sqrt{\sum_{i=1}^n (w_i)^2} \sqrt{\sum_{i=1}^n (v_i)^2}} \quad (7)$$

$$D(w, v) = \frac{2 \sum_{i=1}^n w_i v_i}{\sum_{i=1}^n (w_i)^2 + \sum_{i=1}^n (v_i)^2} \quad (8)$$

$$J(w, v) = \frac{\sum_{i=1}^n w_i v_i}{\sum_{i=1}^n (w_i)^2 + \sum_{i=1}^n (v_i)^2 - \sum_{i=1}^n w_i v_i} \quad (9)$$

$$G(w, v) = \frac{1}{2} \sum_{i=1}^n \left(\frac{\min(w_i, v_i)}{\max(w_i, v_i)} (w_i + v_i) \right) \quad (10)$$

Table 3 presents the scales that are optimal on average in the case of the different discrepancy measures.

As one can see, from the six measures applied in the sensitivity analysis, only two (the Dice similarity, and the Jaccard compatibility indices) provided different optimal scales than the one supported by the Euclidean distance. However, the difference is the smallest possible

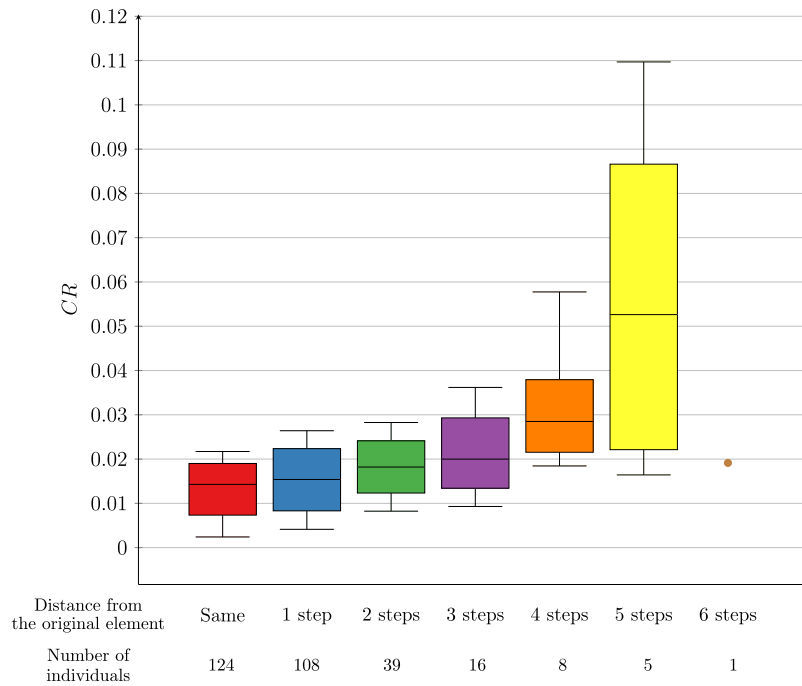


Fig. 5. The CR inconsistency of the individuals depending on the distance (measured in the steps of the verbal scale) of the repeated question's result from the original answer via Box plots (highlighting the minimum, first, second, and third quartile, and the maximum). The number of individuals included in the given category is displayed for each case as well. Instead of a Box plot, only a point representing the corresponding CR is shown for the category with only one individual.

Table 3

The scales that are optimal on average in the case of different discrepancy metrics.

Name	S	M	L
Euclidean distance	1.5	1.7	2
Manhattan distance	1.5	1.7	2
Chebyshev distance	1.5	1.7	2
Cosine similarity	1.5	1.7	2
Dice similarity	1.5	1.7	2.1
Jaccard compatibility	1.5	1.7	2.1
Garuti compatibility	1.5	1.7	2

one on the considered grid, and it only affects one element (L) of the scale. As previously mentioned, the objective function of the Euclidean distance is relatively flat for the problem, i.e., the parameter combinations that are relatively close to the one that is proved to be optimal on average, all provide close-to-optimal values, which remains true for all the examined metrics. Thus, the analysis supports the robustness of the results regarding the chosen discrepancy (similarity) metric. For a comparison of existing scales with all the presented metrics, see [Table A.1](#).

As for measuring the ordinal difference (rank agreement) of the weight vectors calculated from pairwise comparisons and the benchmark direct scoring results, the Spearman and Kendall rank correlation coefficients were considered. However, the scales seem to have quite modest effect on the ranking, as it was also earlier reported in the literature [29]. This was shown by the fact that even in the most conservative case (Kendall rank correlation coefficient and the Eigenvector Method is used) more than 10% of the 236 880 considered parameter combinations provided the same average rank correlations rounded to two decimal places (among which one can find the scale that was optimal according to the Euclidean distance as well). Thus, the flatness of the objective function is even more significant here, and the earlier results of the literature, namely that judgment scales have a significant impact on priorities, but not necessarily on the ranking, is confirmed.

5.2. Sensitivity to the problem instance — further experiments

In order to demonstrate the robustness and generality of the results and corroborate the former analysis, besides repeating the same color-choice test on a different sample of individuals (133 participants), a number of further experiments have been carried out in a sufficiently different setting (e.g., different entities to compare or even different modality stimuli, i.e., smell). The details of the additional tests alongside the original experiments are presented in [Table 4](#).

133 participants' pairwise and direct evaluations were collected and used in a repeated color-choice test, however, afterwards the individuals were also asked the same questions where the colors were presented in reverse order in every pair.

Another 119 evaluators' pairwise comparisons and direct scoring results were collected comparing six completely different entities than before (chocolate, coffee, almond, cinnamon and vanilla). The participants compared the odor of these entities three different ways, based on simple text, then based on an image of each entity, and finally based on the odors of the entities themselves.

The logical structure of the text-based test was identical to that of the original color-choice test, but instead of individual colors, we provided the names of foods in text form, and we tested their odor's likability (see [Figs. A.4](#) and [A.5](#) in [Appendix](#)). The structure of the image-based test is identical to that of the text-based test, but in this test we provided both images and names of the foods (see [Figs. A.6](#) and [A.7](#) in [Appendix](#)). The structure of the odor-based test is identical to that of the text-based and image-based tests (see [Figs. A.8](#) and [A.9](#) in [Appendix](#)). However, we tested the likability of various food odors using cellulose paper strips (150 mm × 7 mm) sealed in odorless, screw-top test tubes, with undiluted aromas absorbed into the bottom 20 mm of the strips. We used food-grade Flavorchem aromas tested during the study (chocolate flavor, coffee natural flavoring, apple flavoring, natural flavoring almond, natural cinnamon flavoring, natural flavoring substances vanilla). During the odor-based test, to minimize sensory fatigue, we took a few-minute break halfway through the pairwise comparison test tasks and before the direct evaluation test tasks.

Table 4
The details of the different experiments.

Name	Entities to compare	Basis of comparison	Number of participants
Original color-choice test	red, green, blue, magenta, turquoise, yellow	Image	301
Repeated color-choice test	red, green, blue, magenta, turquoise, yellow	Image	133
Reversed color-choice test	red, green, blue, magenta, turquoise, yellow	Image	133
Text-based test	chocolate, coffee, apple, almond, cinnamon, vanilla	Text	119
Image-based test	chocolate, coffee, apple, almond, cinnamon, vanilla	Image	119
Odor-based test	chocolate, coffee, apple, almond, cinnamon, vanilla	Odor	119

Table 5
The scales that are optimal on average for each experiment.

Name	<i>S</i>	<i>M</i>	<i>L</i>
Original color-choice test	1.5	1.7	2
Repeated color-choice test	1.4	1.7	2.2
Reversed color-choice test	1.5	1.7	2.1
Text-based test	1.4	1.5	1.8
Image-based test	1.4	1.6	1.8
Odor-based test	1.4	1.5	2

The applied verbal scales and methodological details were the same in these experiments as the ones presented in Section 3, except that for the direct scoring a ten-point scale was used, with 1 indicating a strong dislike and 10 indicating strong preference. This way no further cleaning of the data was necessary.

The scales that result in the smallest average Euclidean distance for both weight calculation techniques in the case of the different experiments are presented in Table 5.

As one can see, although the optimal scales are not always the same, the results are close enough to each other even for completely different experiments, corroborating the former findings.

With this examination, the list of sensitivity analyses is completed, and it is shown that the presented results are robust regarding

- the weight calculation technique (Eigenvector or Logarithmic Least Squares Method);
- the used discrepancy metric (see Section 5.1 and Table 3);
- and the problem instance as well (see Table 5).

6. Conclusion and further research

In this paper, we analyzed the individually and on average optimal numerical scales related to a four-item verbal scale based on a color-choice test carried out on color-calibrated tablets in ISO standardized sensory laboratory and test booths [43]. 462 respondents participated in the experiment providing verbal pairwise comparisons, as well as direct evaluation scores regarding the six colors (red, green, blue, magenta, turquoise, and yellow).

The Eigenvector and the Logarithmic Least Squares Methods were used to calculate the preference representing weight vectors from the pairwise comparison matrices. To compare the obtained weight vectors with the corresponding values calculated from the direct evaluation scores, the Euclidean distance was used. A total of 236 880 different numerical scales were considered.

It turned out that both on average and individually the numerical scales using smaller (≤ 3) numbers even for the extreme preference prevail in this subjective problem. The results do not depend on the used weight calculation technique. This is a significant finding, as it contradicts to both the common practice in Multicriteria Decision Making (especially the AHP), and to several ISO standards related to

sensory tests, where verbal expressions are mechanically converted into great numbers (e.g., 9). Thus, the results suggest that several methods based on pairwise comparisons (for instance, the AHP) and ISO standards need to be revised.

A twofold robustness analysis was carried out to corroborate the findings. First, additional discrepancy metrics were investigated instead of the Euclidean distance, including other distances, rank correlations, and similarity, and compatibility indices. Then a number of further experiments were conducted with different samples of individuals and in some cases with different entities to compare or even with different modality stimuli (i.e., smell). The main findings were the same for all the different metrics and tests. However, the former result from the literature that scales have significant impact on priorities, but modest effect on ranking was confirmed [29].

Following the findings of transforming the verbal expressions of the category scale into numerical values, it is necessary to revisit some of the international standards, such as the ‘Guidelines for the use of quantitative response scales’ [14], ‘Selection and training of sensory assessors’ [52], ‘Guidelines for monitoring the performance of a quantitative sensory panel’ [53], and ‘General guidance for conducting hedonic tests with consumers in a controlled area’ [54].

We also analyzed the inconsistency of the respondents. It was emphasized that the scale itself heavily influences most of the inconsistency measures (for instance the *CR* inconsistency index). The modified random index (*RI*) related to a modified *CR* for the numerical scale that proved to be optimal was calculated. However, the generalization of the known 10% rule is not recommended based on the different number of possibilities (four-item scale instead of nine-item one), the nature (subjective) and size (6×6) of the problem.

A further inconsistency analysis was carried out in the paper, that included a repeated answer to a preference relation between a given pair of colors at the conclusion of the experiment. The ratio of respondents who provided answers that resulted in a sure ordinal difference in the preference (at least a distance of 3 steps in the four-item verbal scale) was a little bit less than 10%. However, $\sim 77\%$ of the respondents answered almost the same way for the second time as well (at most a distance of 1 step in the four-item verbal scale). Examining the inconsistency of respondents whose repeated answer was further from the original, it turned out that they tend to be more inconsistent in general.

Further research questions include analyzing the conversion between verbal and numerical scales on other subjective problems as well, e.g., applying further sensory tests (odor, flavor, etc.); and studying the effect of choosing a different benchmark method instead of direct scoring. A future research can also investigate the extent in which the used numerical scales can be personalized in several fields of decision making and in different practical problems as well.

The use of verbal category scales is widespread in preference studies, so the cognitive conversion of verbal categories into numbers is necessary. Furthermore, it is important to underline that there is currently very limited knowledge available on the stability of individual

Table A.1

Comparison with existing scales using the Eigenvector Method for weight calculation. Each scale is obtained by using the usual choice of parameters presented in Table 1. For the geometric scale, the parameter choice that provides the better results, $\alpha = \sqrt{2}$, is applied. For the affine scale, always the alternative that is first in the (red, green, blue, magenta, turquoise, yellow) list was marked as preferred with 1.1 value, if the decision maker indicated indifference between the given pair. The following notations are applied: d_e –Euclidean distance, d_c –Chebyshev distance, d_m –Manhattan distance, C –cosine similarity index, D –dice similarity index, J –Jaccard compatibility index, G –Garuti compatibility index. For d_e , d_c , and d_m the smaller values show better performance, while in the case of C , D , J , and G , the greater figures are preferred.

Name	d_e	d_c	d_m	C	D	J	G
Empirically optimal scale	0.0800	0.0541	0.1675	0.9816	0.9806	0.9623	0.8546
Linear	0.2245	0.1768	0.4370	0.9035	0.8864	0.8015	0.6642
Affine	0.0855	0.0573	0.1790	0.9785	0.9770	0.9556	0.8469
Power	0.4426	0.3746	0.8399	0.7694	0.7006	0.5547	0.4480
Root	0.0941	0.0654	0.1931	0.9753	0.9744	0.9507	0.8328
Geometric	0.2191	0.1737	0.4242	0.9060	0.8890	0.8074	0.6724
Inverse linear (9/9 – 9/1)	0.0942	0.0661	0.1920	0.9748	0.9737	0.9493	0.8342
Asymptotic	0.0864	0.0595	0.1786	0.9786	0.9778	0.9569	0.8450
Balanced	0.1239	0.0923	0.2457	0.9599	0.9567	0.9188	0.7931
Balanced power	0.1612	0.1240	0.3154	0.9399	0.9325	0.8770	0.7433
Logarithmic	0.1101	0.0785	0.2234	0.9677	0.9660	0.9352	0.8092
Koczkodaj	0.0849	0.0570	0.1777	0.9788	0.9773	0.9562	0.8479

cognitive categories, including their dependence on the type of test item, the modality, the environment, the product being tested, etc. In future research, there will certainly be an increased role for the application and development of numerical conversion methods of verbal category scales.

CRediT authorship contribution statement

Zsombor Szádóczki: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Sándor Bozóki:** Supervision, Conceptualization. **László Sipos:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Zsófia Galambosi:** Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no conflict of interest to report.

Acknowledgments

The project identified by EKOP-CORVINUS-24-4-080 was realized with the support of the National Research, Development, and Innovation Fund provided by the Ministry of Culture and Innovation, as part of the University Research Scholarship Program announced for the 2024/2025 academic year. The research was supported by the National Research, Development and Innovation Office under Grants FK 145838, Advanced 152220, Advanced 150850, Advanced 152184 and TKP2021-NKTA-01 NRDIO. L. Sipos was supported by the Research Excellence Program of the Hungarian University of Agriculture and Life Sciences. Zs. Galambosi was supported by the EKÖP-MATE/2025/26/D university research Scholarship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

Appendix

Which colour do you prefer? *

Click on the picture!




☐ Option 1
 ☒ Option 2

How much more do you like your chosen colour? *

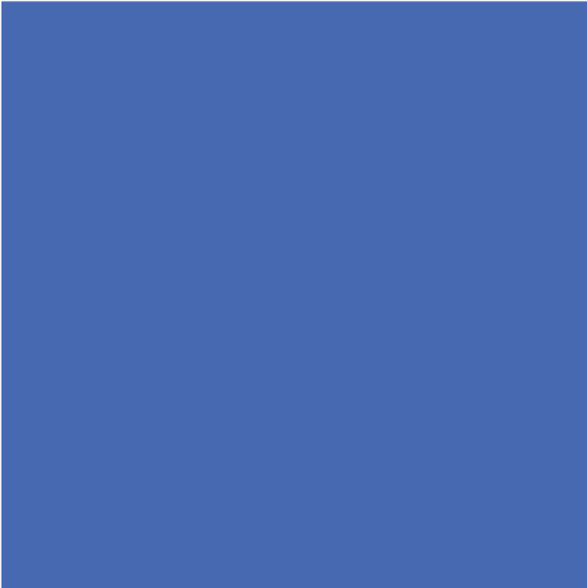
☐ equally like
 ☒ like it a little bit more
 ☐ like it moderately more
 ☐ like it much more

Fig. A.1. A representative example of the questions for pairwise comparisons used in our color-choice test.

How do you like this colour? *

Please rate this colour from 0 to 10!

(0=don't like it at all; 10=favourite colour)



0 1 2 3 4 5 6 7 8 9 10

☐ ☐ ☐ ☐ ☐ ☐ ☒ ☐ ☐ ☐ ☐

Fig. A.2. A representative example of the questions for direct evaluation used in our color-choice test.

How do you like this colour? *

Please rate this colour from 0 to 10!

(0=don't like it at all; 10=favourite colour)



0 1 2 3 4 5 6 7 8 9 10

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☒ ☐ ☐

Fig. A.3. Another example of the questions for direct evaluation used in our color-choice test.

Which food odor do you prefer? *

Click on the picture!

chocolate

almond

☐ chocolate

☒ almond

How much more do you like your chosen food odor? *

☐ equally like

☐ like it a little bit more

☒ like it moderately more

☐ like it much more

Fig. A.4. A representative example of the questions for pairwise comparisons used in the text-based test.

How do you like the chocolate odor? *

Please rate the chocolate odor from 0 to 10!

(0=don't like it at all; 10=favourite odor)

chocolate

1

2

3

4

5

6

7

8

9

10

☐☐☐☐☐☐☐☒☐☐

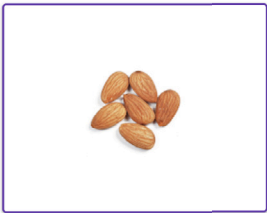
Fig. A.5. A representative example of the questions for direct evaluation used in the text-based test.

Which food odor do you prefer? *

Click on the picture!



☐ chocolate



☒ almond

How much more do you like your chosen food odor? *

☐ equally like
☐ like it a little bit more
☒ like it moderately more
☐ like it much more

Fig. A.6. A representative example of the questions for pairwise comparisons used in the image-based test.

How do you like the chocolate odor? *

Please rate the chocolate odor from 0 to 10!
(0=don't like it at all; 10=favourite odor)



12345678910

☐☐☐☐☐☐☐☒☐☐

Fig. A.7. A representative example of the questions for direct evaluation used in the image-based test.

Which food odor do you prefer? *

Click on the picture!

chocolate

☐ chocolate

almond

☒ almond

How much more do you like your chosen food odor? *

☐ equally like
☐ like it a little bit more
☒ like it moderately more
☐ like it much more

Fig. A.8. A representative example of the questions for pairwise comparisons used in the odor-based test.

How do you like the chocolate odor? *

Please rate the chocolate odor from 0 to 10!
(0=don't like it at all; 10=favourite odor)

chocolate

12345678910

☐☐☐☐☐☐☐☒☐☐

Fig. A.9. A representative example of the questions for direct evaluation used in the odor-based test.

Data availability

Data will be made available on request.

References

- [1] Triantaphyllou E. Multi-criteria decision making methods. In: Multi-criteria Decision Making Methods: A Comparative Study. Applied Optimization, vol 44, Boston, MA: Springer; 2000. http://dx.doi.org/10.1007/978-1-4757-3157-6_2.
- [2] Ishizaka A, Labib A. Review of the main developments in the analytic hierarchy process. Expert Syst Appl 2011;38(11):14336–45. <http://dx.doi.org/10.1016/j.eswa.2011.04.143>.
- [3] Greco S, Słowiński R, Wallenius J. Fifty years of multiple criteria decision analysis: From classical methods to robust ordinal regression. European J Oper Res 2025;323(2):351–77. <http://dx.doi.org/10.1016/j.ejor.2024.07.038>.
- [4] Davidson R, Farquhar P. A bibliography on the method of paired comparisons. Biometrics 1976;32(2):241–52.
- [5] Al-Husain R, Khorramshahgol R. Incorporating analytical hierarchy process and goal programming to design responsive and efficient supply chains. Oper Res Perspect 2020;7:100149. <http://dx.doi.org/10.1016/j.orp.2020.100149>.
- [6] Csató L. Tournament design: how operations research can improve sports rules. In: Palgrave Pivots in Sports Economics. 2021. <http://dx.doi.org/10.1007/978-3-030-59844-0>.
- [7] Saaty TL. A scaling method for priorities in hierarchical structures. J Math Psych 1977;15(3):234–81. [http://dx.doi.org/10.1016/0022-2496\(77\)90033-5](http://dx.doi.org/10.1016/0022-2496(77)90033-5).
- [8] Saaty TL. The Analytic Hierarchy Process. New York: McGraw-Hill; 1980.
- [9] Bozóki S, Dezső L, Poesz A, Temesi J. Analysis of pairwise comparison matrices: an empirical research. Ann Oper Res 2013;211(1):511–28. <http://dx.doi.org/10.1007/s10479-013-1328-1>.
- [10] Cavallo B, Ishizaka A. Evaluating scales for pairwise comparisons. Ann Oper Res 2023;325(2):951–65. <http://dx.doi.org/10.1007/s10479-022-04682-8>.
- [11] Szádóczi Zs, Bozóki S, Tekile HA. Filling in pattern designs for incomplete pairwise comparison matrices: (quasi-)regular graphs with minimal diameter. Omega 2022;107:102557. <http://dx.doi.org/10.1016/j.omega.2021.102557>.
- [12] Simonson I. Will I like a “medium” pillow? another look at constructed and inherent preferences. J Consum Psychol 2008;18(3):155–69. <http://dx.doi.org/10.1016/j.jcps.2008.04.002>.
- [13] Cavallo B, Ishizaka A. A comparative study on precision of direct evaluations, the pairwise comparisons method and the best-worst method. Omega 2025;130:103175. <http://dx.doi.org/10.1016/j.omega.2024.103175>.
- [14] ISO, 4121 International Organization for Standardization. Sensory analysis – guidelines for the use of quantitative response scales. 2003.
- [15] Rezaei J. Best-worst multi-criteria decision-making method. Omega 2015;53:49–57. <http://dx.doi.org/10.1016/j.omega.2014.11.009>.
- [16] Huizingh EK, Vrolijk HC. A comparison of verbal and numerical judgments in the analytic hierarchy process. Organ Behav Hum Decis Process 1997;70(3):237–47. <http://dx.doi.org/10.1006/obhd.1997.2708>.
- [17] Pöyhönen MA, Hämäläinen RP, Salo AA. An experiment on the numerical modelling of verbal ratio statements. J Multi-Criteria Decis Anal

- 1997;6(1):1–10. [http://dx.doi.org/10.1002/\(SICI\)1099-1360\(199701\)6:1<1::AID-MCDA111>3.0.CO;2-W](http://dx.doi.org/10.1002/(SICI)1099-1360(199701)6:1<1::AID-MCDA111>3.0.CO;2-W).
- [18] Salo AA. Inconsistency analysis by approximately specified priorities. *Math Comput Modelling* 1993;17(4):123–33. [http://dx.doi.org/10.1016/0895-7177\(93\)90181-W](http://dx.doi.org/10.1016/0895-7177(93)90181-W).
- [19] Salo AA, Hämäläinen RP. Preference programming through approximate ratio comparisons. *European J Oper Res* 1995;82(3):458–75. [http://dx.doi.org/10.1016/0377-2217\(93\)E0224-L](http://dx.doi.org/10.1016/0377-2217(93)E0224-L).
- [20] Elliott MA. Selecting numerical scales for pairwise comparisons. *Reliab Eng Syst Saf* 2010;95(7):750–63. <http://dx.doi.org/10.1016/j.res.2010.02.013>.
- [21] Finan J, Hurley W. Transitive calibration of the AHP verbal scale. *European J Oper Res* 1999;112(2):367–72. [http://dx.doi.org/10.1016/S0377-2217\(97\)00411-6](http://dx.doi.org/10.1016/S0377-2217(97)00411-6).
- [22] Dong Y, Hong W-C, Xu Y, Yu S. Numerical scales generated individually for analytic hierarchy process. *European J Oper Res* 2013;229(3):654–62. <http://dx.doi.org/10.1016/j.ejor.2013.03.019>.
- [23] Rokou E, Kirytopoulos K. A calibrated group decision process. *Group Decis Negot* 2014;23(6):1369–84. <http://dx.doi.org/10.1007/s10726-013-9374-2>.
- [24] Meesariganda BR, Ishizaka A. Mapping verbal AHP scale to numerical scale for cloud computing strategy selection. *Appl Soft Comput* 2017;53:111–8. <http://dx.doi.org/10.1016/j.asoc.2016.12.040>.
- [25] Ahmed F, Kilic K. A basic algorithm for generating an individualized numerical scale. *Expert Syst Appl* 2023;233:120915. <http://dx.doi.org/10.1016/j.eswa.2023.120915>.
- [26] Dutta B, Labella Á, Ishizaka A, Martínez L. Eliciting personalized AHP scale from verbal pairwise comparisons. *J Oper Res Soc* 2025;76(3):541–53. <http://dx.doi.org/10.1080/01605682.2024.2376033>.
- [27] Koczkodaj WW, Kakiashvili T, Szymańska A, Montero-Marin J, Araya R, Garcia-Campayo J, Rutkowski K, Strzałka D. How to reduce the number of rating scale items without predictability loss? *Scientometrics* 2017;111(2):581–93. <http://dx.doi.org/10.1007/s11192-017-2283-4>.
- [28] Koczkodaj WW, Kakiashvili T, Li F, Wolny-Dominiak A, Masiak J. Optimizing predictability of rating scales by differential evolution for the use by collective intelligent information and database systems. *J Intell Fuzzy Systems* 2019;37(6):7545–53. <http://dx.doi.org/10.3233/JIFS-179361>.
- [29] Franek J, Kresta A. Judgment scales and consistency measure in AHP. In: *Procedia Economics and Finance*. 12, 2014, p. 164–73. [http://dx.doi.org/10.1016/S2212-5671\(14\)00332-3](http://dx.doi.org/10.1016/S2212-5671(14)00332-3), 17th international conference enterprise and competitive environment 2014.
- [30] Goepel KD. Comparison of judgment scales of the analytical hierarchy process — a new approach. *Int J Inf Technol Decis Mak* 2019;18(02):445–63. <http://dx.doi.org/10.1142/S0219622019500044>.
- [31] Saaty R. The analytic hierarchy process—what it is and how it is used. *Math Model* 1987;9(3):161–76. [http://dx.doi.org/10.1016/0270-0255\(87\)90473-8](http://dx.doi.org/10.1016/0270-0255(87)90473-8).
- [32] Harker PT, Vargas LG. The theory of ratio scale estimation: Saaty's analytic hierarchy process. *Manag Sci* 1987;33(11):1383–403. <http://dx.doi.org/10.1287/mnsc.33.11.1383>.
- [33] Lootsma F. Conflict resolution via pairwise comparison of concessions. *European J Oper Res* 1989;40(1):109–16. [http://dx.doi.org/10.1016/0377-2217\(89\)90278-6](http://dx.doi.org/10.1016/0377-2217(89)90278-6).
- [34] Ma D, Zheng X. 9/9–9/1 scale method of AHP. In: *Proceedings of the 2nd international symposium on the AHP*. vol. 1, 1991, p. 197–202.
- [35] Donegan HA, Dodd FJ, McMaster TBM. A new approach to AHP decision-making. *J R Stat Soc Ser D (the Statistician)* 1992;41(3):295–302. <http://dx.doi.org/10.2307/2348551>.
- [36] Salo AA, Hämäläinen RP. On the measurement of preferences in the analytic hierarchy process. *J Multi-Criteria Decis Anal* 1997;6(6):309–19. [http://dx.doi.org/10.1002/\(SICI\)1099-1360\(199711\)6:6<309::AID-MCDA163>3.0.CO;2-2](http://dx.doi.org/10.1002/(SICI)1099-1360(199711)6:6<309::AID-MCDA163>3.0.CO;2-2).
- [37] Ishizaka A, Balkenborg D, Kaplan T. Influence of aggregation and measurement scale on ranking a compromise alternative in AHP. *J Oper Res Soc* 2011;62(4):700–10. <http://dx.doi.org/10.1057/jors.2010.23>.
- [38] Koczkodaj WW. Pairwise comparisons rating scale paradox. In: Nguyen NT, Kowalczyk R, editors. *Transactions on Computational Collective Intelligence XXII*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016, p. 1–9. http://dx.doi.org/10.1007/978-3-662-49619-0_1.
- [39] Brunelli M. A survey of inconsistency indices for pairwise comparisons. *Int J Gen Syst* 2018;47(8):751–71. <http://dx.doi.org/10.1080/03081079.2018.1523156>.
- [40] Brunelli M, Fedrizzi M. Inconsistency indices for pairwise comparisons and the Pareto dominance principle. *European J Oper Res* 2024;312(1):273–82. <http://dx.doi.org/10.1016/j.ejor.2023.06.033>.
- [41] Crawford G, Williams C. A note on the analysis of subjective judgment matrices. *J Math Psych* 1985;29(4):387–405. [http://dx.doi.org/10.1016/0022-2496\(85\)90002-1](http://dx.doi.org/10.1016/0022-2496(85)90002-1).
- [42] Mylonas D, MacDonald L. 5 - colour naming for colour design, In: Best J, editor. *In: Woodhead Publishing Series in Textiles, Colour Design (Second Edition)*. Woodhead Publishing; 2017, p. 131–55. <http://dx.doi.org/10.1016/B978-0-08-101270-3.00005-9>, second ed..
- [43] ISO, 8589 International Organization for Standardization. *Sensory analysis – general guidance for the design of test rooms*. 2007.
- [44] Gass SI, Standard SM. Characteristics of positive reciprocal matrices in the analytic hierarchy process. *J Oper Res Soc* 2002;53(12):1385–9. <http://dx.doi.org/10.1057/palgrave.jors.2601471>.
- [45] Poesz A. Inconsistency in Multi-Attribute Decision Problems. (Ph.D. thesis), Corvinus University of Budapest. In Hungarian; 2022. <http://dx.doi.org/10.14267/phd.2022053>.
- [46] Bozóki S, Rapcsák T. On saaty's and koczkodaj's inconsistencies of pairwise comparison matrices. *J Global Optim* 2008;42(2):157–75. <http://dx.doi.org/10.1007/s10898-007-9236-z>.
- [47] Ágoston KCs, Csató L. Inconsistency thresholds for incomplete pairwise comparison matrices. *Omega* 2022;108:102576. <http://dx.doi.org/10.1016/j.omega.2021.102576>.
- [48] Kou G, Peng Y, Chao X, Herrera-Viedma E, Alsaadi FE. A geometrical method for consensus building in GDM with incomplete heterogeneous preference information. *Appl Soft Comput* 2021;105:107224. <http://dx.doi.org/10.1016/j.asoc.2021.107224>.
- [49] Ye J. Multicriteria decision-making method using the dice similarity measure based on the reduct intuitionistic fuzzy sets of interval-valued intuitionistic fuzzy sets. *Appl Math Model* 2012;36(9):4466–72. <http://dx.doi.org/10.1016/j.apm.2011.11.075>.
- [50] Jaccard P. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull de la Société Vaudoise Des Sci Nat* 1901;41:241–72, (in French).
- [51] Garuti CE. A set theory justification of Garuti's compatibility index. *J Multi-Criteria Decis Anal* 2020;27(1–2):50–60. <http://dx.doi.org/10.1002/mcda.1667>.
- [52] ISO, 8586 International Organization for Standardization. *Sensory analysis – selection and training of sensory assessors*. 2023.
- [53] ISO, 11132 International Organization for Standardization. *Sensory analysis – methodology – guidelines for monitoring the performance of a quantitative sensory panel*. 2012.
- [54] ISO, 11136 International Organization for Standardization. *Sensory analysis – methodology – general guidance for conducting hedonic tests with consumers in a controlled area*. 2014.