Annals of Operations Research (2023) 325:951–965
https://doi.org/10.1007/s10479-022-04682-8

ORIGINAL RESEARCH

Evaluating scales for pairwise comparisons

Bice Cavallo[1] · **Alessio Ishizaka**[2]

Accepted: 18 March 2022 / Published online: 7 April 2022

Abstract
Pairwise comparisons have been a long-standing technique for comparing alternatives/criteria and their role has been pivotal in the development of modern decision-making methods. The evaluation is very often done linguistically. Several scales have been proposed to translate the linguistic evaluation into a quantitative evaluation. In this paper, we perform an experiment to investigate, under our methodological choices, which type of scale provides the best matching of the decision-maker's verbal representation. The experiment aims to evaluate the suitability of eight evaluation scales for problems of different sizes. We find that the inverse linear scale provides the best matching verbal representation whenever the objective data are measured by means of pairwise comparisons matrices and a suitable distance between matrices is applied for computing the matching error.

Keywords Decision analysis · Multi-criteria · Pairwise comparison matrix · Evaluation scales

1 Introduction

Being able to make a good decision is fundamental for the survival of companies, people and societies. However, to make good decisions, we need to be able to measure performance. If a performance is not directly objectively measurable, then it can be evaluated. Subjective evaluations are often given in a linguistic scale (e.g., poor, good, excellent, etc) with a better result than numeric evaluations (Windschitl and Wells 1996). The question is how to convert a verbal scale into a numerical scale that can be used in the prioritization of actions. Although several scales have been proposed (Meesariganda and Ishizaka 2017), it is not clear which one to choose; indeed, the authors have seen that there is not a unique scale for the conversion of a fixed problem (i.e., cloud computing strategy selection): each person has its own representation. To investigate this question, we propose an experimental method to evaluate the same

✉ Bice Cavallo
bice.cavallo@unina.it

Alessio Ishizaka
alessio.ishizaka@neoma-bs.fr

[1] Department of Architecture, University of Naples Federico II, Via Toledo 402, 80134 Naples, Italy

[2] Information Systems, Supply Chain and Decision making Department, NEOMA Business School, 1 rue du Maréchal Juin - BP 215, 76130 Mont-Saint-Aignan Cedex, France

scales used by Meesariganda and Ishizaka (2017), but in further problems. Experiments are very popular in psychology (Wixted 2018). For example, experimental psychology employs human participants to study a variety of topics, including among others sensation and perception, memory, cognition, learning, motivation and emotion. Experiments have been also used in economics (Kagel and Roth 2017) and in operational research (Donohue et al. 2018). An experiment was performed by Keeney et al. (1990), who asked participants to provide a direct ranking of options. They then solved the problem with the Multi Attribute Utility (MAU). In the final ranking, 80% of the participants changed their initial ranking, mostly in agreement with the MAU ranking. Huizingh and Vrolijk (1997) asked participants to select a room to rent. They observed that the participants were more satisfied with the AHP result than with a random selection. Linares (2009) asked 18 participants to rank cars with AHP. Then, an automatic algorithm removed the intransitivities and a new ranking was generated. In a questionnaire, the majority of the students said that when intransitivities were removed, their preferences were not better represented. Ishizaka et al. (2011) statistically compared three rankings: one given directly by the participants, a second by applying AHP, a third after having provided the information for AHP and a final after learning the ranking provided by AHP. Their results showed that the rankings were similar. Furthermore, it was found that if participants changed their ranking, then they followed the suggestion of AHP. Later, Ishizaka and Siraj (2018) renewed the experiment with a different decision problem solved with AHP, MACBETH and SMART, and the same results were found. Bozóki et al. (2013) conducted experiments for testing various characteristics of pairwise comparison matrices. Recently, Cavallo et al. (2019) measured the coherence of the participants when they express their subjective preferences by means of additive, multiplicative and fuzzy pairwise comparisons. They found that the worst level of coherence occurs when participants use the additive pairwise comparison. An algorithm for controlling the incoherence during the construction of the pairwise comparison matrix has been proposed by Ishizaka and Siraj (2020). Relations among several coherence conditions have been provided by Brunelli and Cavallo (2020b) and Cavallo and D'Apuzzo (2020).

When we have to solve a problem, we may do not know in advance the range of values in the problem. The goal of this paper is to investigate the existence of a scale that always provides the best matching verbal representation of the respondents; therefore, we have selected problems with different ranges to analyze whether a scale would always perform better whatever the range.

In this paper, we ask participants to evaluate three problems of different known dimensions with a verbal scale. Then, the verbal evaluations are matched with the real values of eight numerical scales. We find that, under our methodological choices, the inverse linear scale provides the best matching verbal representation of the respondents, followed shortly by the root square scale, while the power scale is the worst.

The remainder of this paper is organized as follows. Section 2 presents the different evaluation scales and provides basic theoretical notions about pairwise comparison matrices. The experimental part and its results are given in Sect. 3. Finally, Sect. 4 concludes this paper and outlines some future avenues of research.

2 Preliminaries

In this section, we provide preliminaries useful in the sequel.

2.1 Evaluation scales

Pairwise comparisons have been used in psychology since the beginning of last century (Yokoyama 1921), (Thurstone 1927). They have also been adopted in multi-criteria decision analysis, for example in AHP (Saaty 1977) and BWM (Rezaei 2015). Decision items are compared in pairs and their evaluation is entered into a squared matrix. The priorities are then calculated from this matrix. It has previously been experimentally demonstrated that pairwise comparisons are more precise than direct evaluations (Millet 1997), (Por and Budescu 2017), (Whitaker 2007). Generally, the comparisons are expressed on a verbal scale because it is more familiar to decision-makers and they understand it better than numerical evaluations (Budescu and Wallsten 1985). The verbal scale is only converted into numerical values to calculate priorities in a second step without the decision-maker. Several scale conversions have been proposed. The most commonly used is the linear scale in Table 1 that was proposed by Saaty, probably because it is integrated into the leading Expert choice software program (Ishizaka and Labib 2009).

It has been demonstrated that different decision-makers may have different interpretations of the same verbal expression (Huizingh and Vrolijk 1997). Therefore, several scales have been proposed (Table 2). For example, another linear scale was proposed by Ma and Zheng

Table 1 Saaty's scale

Value	Semantic meaning
1	The two decision items have the same importance
3	The selected decision item is weakly more important than the other one
5	The selected decision item is strongly more important than the other one
7	The selected decision item is very strongly more important than the other one
9	The selected decision item is absolutely more important than the other one
2, 4, 6, 8	Intermediate values

Table 2 Evaluation scales

Scale type	Definition	Parameters
1. Linear	$c = \alpha \cdot x$	$\alpha > 0, \quad x \in \{1,\ 2, \ldots, 9\}$
2. Power	$c = x^{\alpha}$	$\alpha > 1, \quad x \in \{1,\ 2, \ldots, 9\}$
3. Geometric	$c = \alpha^{x-1}$	$\alpha > 1,$ $x \in \{1,\ 2, \ldots, 9\}$ or $x \in \{1,\ 1.15,\ \ldots, 4\}$ or other step
4. Logarithmic	$c = log_{\alpha}(x + (\alpha - 1))$	$\alpha > 1, \quad x \in \{1,\ 2, \ldots, 9\}$
5. Root square	$c = \sqrt[\alpha]{x}$	$\alpha > 1, \quad x \in \{1,\ 2, \ldots, 9\}$
6. Inverse linear	$c = \frac{9}{10-x}$	$x \in \{1,\ 2, \ldots, 9\}$
7. Balanced	$c = \frac{x}{1-x}$	$x \in \{0.5,\ 0.55, \ldots, 0.9\}$
8. Balanced power	$c = 9^{\frac{x-1}{n-1}}$	$x \in \{1,\ 2, \ldots, n\}$

(1991), where the inverse elements (i.e. $\frac{1}{x}$) are linear instead of the x that is used in the Saaty's scale.

The power and root scale were proposed by Harker and Vargas (1987), but they then recognised that the linear scale would often outperform them. Inspired from auditory stimuli, the geometric scale was proposed to represent equidistant stimuli (Lootsma 1989). Salo and Hämäläinen (1997) observed that when two alternatives are compared on an integer scale one to nine, there is an uneven dispersion of local priorities. This means that there is a lack of sensitivity when comparing elements, which are preferentially close to each other. Based on this observation, the authors proposed a balanced scale where the local priorities are evenly dispersed between the range [0.1, 0.9]. Later, Elliott (2010) developed a balanced power scale. This scale is based on the same idea as Salo and Hämäläinen (1997) but with balanced priorities among three alternatives. To avoid the boundary problem that is incurred by the consistency rule (e.g., if the decision-maker enters $a_{ij} = 4$ and $a_{jk} = 5$, the user is forced to an inconsistent relation when the upper limit of the scale is 9, and therefore cannot enter $a_{ik} = 20$), Donegan et al. (1992) proposed an asymptotic scale. Ishizaka et al. (2011) observed that the existing scales disadvantage compromised solutions. To solve this problem, they suggested a logarithmic scale, which provides a better priority for compromised solutions.

Although these scales have advantages and disadvantages, the only question that counts is: what is the scale in the mind of the decision-maker? To solve this problem, individual scales have been developed. Dong et al. (2013) and Liang et al. (2008) proposed a method to calculate personalised scales that minimise the inconsistency of the matrix. However, these techniques are questionable because inconsistency in the pairwise matrix can have several origins in addition to the scale effect, including error, lack of information, distracted or undecided user, and so on. Another approach to choose the most appropriate scale has been to map verbal scale with comparisons given by the decision-maker on alternatives with known measures, such as the surface of geometrical figures. This technique has been used for fuzzy AHP (Ishizaka and Nguyen 2013), ANP (Rokou and Kirytopoulos 2014) and AHP (Meesariganda and Ishizaka 2017). However, the chosen scale may depend on the objective problem that is chosen to calibrate the scale. Scale invariance property has been the subject of some recent papers (Csató 2017, 2018, 2019).

2.2 Pairwise comparison matrices

In this section, we provide the theoretical framework about Pairwise Comparison Matrices (PCMs) that we choose for the experiment; we adopt the multiplicative representation (Saaty 1977). Thus, from now on, we assume that each PCM satisfies the following reciprocity property:

Definition 2.1 $\mathbf{M} = [m_{ij}]_{n\times n}$, with $m_{ij} > 0$, is a *reciprocal* PCM if, for all $i, j \in \{1, \ldots, n\}$, it verifies the following condition:

$$m_{ji} = \frac{1}{m_{ij}}.$$

The distance (Cavallo 2019) between $\mathbf{M}^1 = [m^1_{ij}]_{n\times n}$ and $\mathbf{M}^2 = [m^2_{ij}]_{n\times n}$ is:

$$d(\mathbf{M^1}, \mathbf{M}^2) = \sqrt[\frac{n(n-1)}{2}]{\prod_{i=1}^{n-1}\prod_{j=i+1}^{n} d(m^1_{ij}, m^2_{ij})}; \tag{1}$$

Table 3 Evaluation scales used in the experiment

	a	b	c	d	e	f	g	h	i
1. Linear ($\alpha = 1$)	1	2	3	4	5	6	7	8	9
2. Power ($\alpha = 2$)	1	4	9	16	25	36	49	64	81
3. Geometric ($\alpha = 2$)	1	2	4	8	16	32	64	128	256
4. Logarithmic ($\alpha = 2$)	1	1.58	2	2.32	2.58	2.81	3	3.17	3.32
5. Root square ($\alpha = 2$)	1	1.41	1.73	2	2.24	2.45	2.65	2.83	3
6. Inverse linear	1	1.13	1.29	1.5	1.8	2.25	3	4.5	9
7. Balanced	1	1.22	1.5	1.86	2.33	3	4	5.67	9
8. Balanced power ($n = 9$)	1	1.32	1.73	2.28	3	3.95	5.2	6.84	9

where

$$d(m^1_{ij}, m^2_{ij}) = max\{m^1_{ij}/m^2_{ij}, m^2_{ij}/m^1_{ij}\} \tag{2}$$

is the distance between the entries m^1_{ij} and m^2_{ij} (Cavallo and D'Apuzzo 2009). Thus, the distance between $\mathbf{M}^1 = [m^1_{ij}]_{n\times n}$ and $\mathbf{M}^2 = [m^2_{ij}]_{n\times n}$ is the geometric mean of the component-wise distances in the upper triangles of $\mathbf{M}^1 = [m^1_{ij}]_{n\times n}$ and $\mathbf{M}^2 = [m^2_{ij}]_{n\times n}$.

It is nowadays widely acknowledged that seemingly different types of pairwise comparisons (e.g. additive (Barzilai 1998), multiplicative and fuzzy (Tanino 1984)) are mathematically equivalent (Cavallo and D'Apuzzo 2009; Ramík 2015) because they share the same algebraic structure; that is, they are pairwise comparisons defined over Abelian linearly ordered (Alo)-groups. It has been shown that, by means of proper isomorphisms, one can naturally extend concepts and properties from one representation to another. As an example, for two additive PCMs $\mathbf{A}^1 = [a^1_{ij}]_{n\times n}$ and $\mathbf{A}^2 = [a^2_{ij}]_{n\times n}$, the distance equivalent to (1) is the following one:

$$d(\mathbf{A^1}, \mathbf{A}^2) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |a^1_{ij} - a^2_{ij}|; \tag{3}$$

that is, the arithmetic mean of the component-wise absolute differences in the upper triangles of $\mathbf{A}^1 = [a^1_{ij}]_{n\times n}$ and $\mathbf{A}^2 = [a^2_{ij}]_{n\times n}$.

Because of its extensibility, e.g. to the additive and fuzzy PCMs, in our experiment we will use distance in (1) between multiplicative PCMs, in order to evaluate the scales in Table 3. Moreover, remaining in the same theoretical framework, the same distance could be also used, in a future work, for measuring several levels of coherence (Brunelli and Cavallo 2020a, b) of the PCMs in the experiment.

3 The experiment

Our experiment aims to evaluate the suitability of the evaluation scales in Table 3, that are the same scales used by Meesariganda and Ishizaka (2017), for problems of different sizes with known measures. For this purpose, we compare the scales against known measurable evaluations (i.e., distances between cities). In the next section, we will describe the problems in more detail.

Table 4 Verbal scale for measuring how much farther the selected city is from Naples than the other city

Value	Semantic meaning
a	The two cities have the same distance from Naples
c	The selected city is weakly farther from Naples than the other one
e	The selected city is strongly farther from Naples than the other one
g	The selected city is very strongly farther from Naples than the other one
i	The selected city is absolutely farther from Naples than the other one
b, d, f h	Intermediate values

3.1 Problem description

We deal with the following three problems:

P_A. Evaluating distance ratios between cities in Campania region;
P_B. Evaluating distance ratios between cities in Italy;
P_C. Evaluating distance ratios between cities in Europe.

Thus, given the following sets:

$$X_A = \{Caserta,\ Avellino,\ Salerno, Benevento\},$$
$$X_B = \{Rome,\ Florence,\ Bologna,\ Milano,\ Bolzano\},$$
$$X_C = \{Rome,\ Paris,\ Berlin,\ Oslo,\ Moscow\},$$

and the corresponding distances, expressed in km, between each city and Naples, that is:

$$A = \{a_1,\ a_2,\ a_3, a_4\} = \{25,\ 45,\ 45,\ 53\},$$
$$B = \{b_1,\ b_2,\ b_3, b_4,\ b_5\} = \{191,\ 409,\ 471,\ 659,\ 671\},$$
$$C = \{c_1,\ c_2,\ c_3, c_4,\ c_5\} = \{191,\ 1293,\ 1301,\ 2137,\ 2374\},$$

for each pair in X_A, X_B and X_C, a decision maker has to select the farthest city from Naples, and to measure how much farther the selected city is from Naples than the other city; they have to express a verbal judgement among those given in Table 4.

The real distance ratios among the cities in X_A, X_B and X_C are represented by the following multiplicative PCMs:

$$\mathbf{A} = [a_{ij}]_{4\times4} = \left[\frac{a_i}{a_j}\right]_{4\times4} = \begin{bmatrix} 1 & 0.556 & 0.556 & 0.472 \\ 1.8 & 1 & 1 & 0.849 \\ 1.8 & 1 & 1 & 0.849 \\ 2.120 & 1.178 & 1.178 & 1 \end{bmatrix}; \tag{4}$$

$$\mathbf{B} = [b_{ij}]_{5\times5} = \left[\frac{b_i}{b_j}\right]_{5\times5} = \begin{bmatrix} 1 & 0.467 & 0.406 & 0.290 & 0.285 \\ 2.141 & 1 & 0.868 & 0.621 & 0.610 \\ 2.466 & 1.152 & 1 & 0.715 & 0.702 \\ 3.450 & 1.611 & 1.399 & 1 & 0.982 \\ 3.513 & 1.641 & 1.425 & 1.018 & 1 \end{bmatrix}; \tag{5}$$

$$\mathbf{C} = [c_{ij}]_{5\times5} = \left[\frac{c_i}{c_j}\right]_{5\times5} = \begin{bmatrix} 1 & 0.148 & 0.147 & 0.089 & 0.080 \\ 6.770 & 1 & 0.994 & 0.605 & 0.545 \\ 6.812 & 1.006 & 1.000 & 0.609 & 0.548 \\ 11.188 & 1.653 & 1.643 & 1 & 0.900 \\ 12.429 & 1.836 & 1.825 & 1.111 & 1 \end{bmatrix}; \tag{6}$$

respectively. Our experiment aims to assess the suitability of the scales in Table 3 to represent the entries of the PCMs $\mathbf{A} = [a_{ij}]_{4\times4}$, $\mathbf{B} = [b_{ij}]_{5\times5}$ and $\mathbf{C} = [c_{ij}]_{5\times5}$.

3.2 Methodology

We use an opinion survey with a sample of 66 students of the Department of Architecture of University of Naples Federico II, Italy. The students received instructions on how to fill out the survey and they completed it during a class period. On average, the experiment lasted 20 minutes. The survey forms are as follows:

- $\boldsymbol{Q_A}$ The respondent has to perform four pairwise comparisons. For each pairwise comparison of cities in X_A, they have to select the farthest from Naples and they also have to select a verbal judgement from Table 4 to express how much farther the selected city is from Naples than the other city;
- $\boldsymbol{Q_B}$ The respondent has to perform five pairwise comparisons. For each pairwise comparison of cities in X_B, they have to select the farthest from Naples and they have to select a verbal judgement from Table 4 to express how much farther the selected city is from Naples than the other city;
- $\boldsymbol{Q_C}$ The respondent has to perform five pairwise comparisons. For each pairwise comparison of cities in X_C, they have to select the farthest from Naples and they also have to select a verbal judgement from Table 4 to express how much farther the selected city is from Naples than the other city.

The survey forms have been developed by means of Google Forms in such a way that each question is mandatory. For example, Fig. 1 provides a question in the survey Q_A; as we can see, the participants have full information: the map with the cities and the distance to Naples.

We discard surveys where the students select a wrong city as the farthest from Naples because we assume a minimal level of coherence; thus, let m be the number of students who always select the correct city. Although discarding all surveys of a student committing a mistake in any survey Q_A, Q_B or Q_C is selective, this choice ensures that, for each problem, we have the same number of surveys to analyze and that, for each student, we can obtain the type of scale that provides the best matching of his/her verbal representation for each problem.

To evaluate the suitability of the scales in Table 3 for the given problems, we perform the following steps:

Step 1 For each student (i.e., for each $k \in \{1, \ldots, m\}$), the verbal judgements provided in the surveys Q_A, Q_B and Q_C are matched with the real values of the scales in Table 3 by building the PCMs $\mathbf{A}^{ks} = [a_{ij}^{ks}]_{4\times4}$, $\mathbf{B}^{ks} = [b_{ij}^{ks}]_{5\times5}$ and $\mathbf{C}^{ks} = [c_{ij}^{ks}]_{5\times5}$, where $s = \{1, \ldots, 8\}$ is the index of the scale in Table 3 (e.g., $s = 2$ refers to the power scale);

Step 2 For each student (i.e. for each $k \in \{1, \ldots, m\}$) and for each scale (i.e., for each $s \in \{1, \ldots, 8\}$), by applying (1), we compute the following matching errors:

$$d(\mathbf{A}, \mathbf{A}^{ks}) = \sqrt[6]{\prod_{i=1}^{3}\prod_{j=i+1}^{4} \max\{a_{ij}/a_{ij}^{ks}, a_{ij}^{ks}/a_{ij}\}}; \tag{7}$$

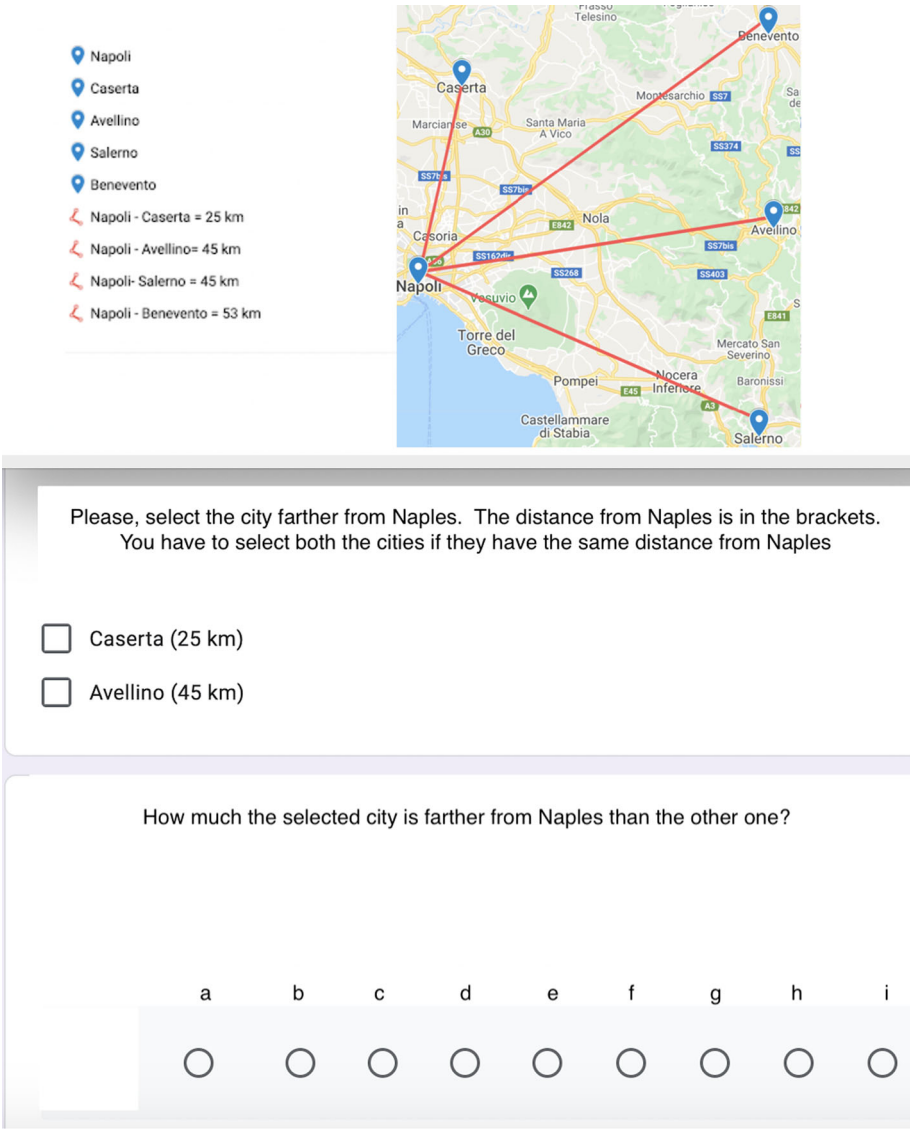


Fig. 1 Example of question in the survey Q_A

$$d(\mathbf{B}, \mathbf{B}^{ks}) = \sqrt[10]{\prod_{i=1}^{4} \prod_{j=i+1}^{5} \max\{b_{ij}/b_{ij}^{ks}, b_{ij}^{ks}/b_{ij}\}}; \tag{8}$$

$$d(\mathbf{C}, \mathbf{C}^{ks}) = \sqrt[10]{\prod_{i=1}^{4} \prod_{j=i+1}^{5} \max\{c_{ij}/c_{ij}^{ks}, c_{ij}^{ks}/c_{ij}\}}. \tag{9}$$

Example 3.1 Let us suppose that student k provides the following verbal judgements to the survey Q_A:

1. Avellino is **weakly farther** (i.e., value= **c** in Table 4) from Naples than Caserta;
2. Salerno is **weakly farther** (i.e., value= **c** in Table 4) from Naples than Caserta;
3. Benevento is **weakly farther** (i.e., value= **c** in Table 4) from Naples than Caserta;
4. Avellino and Salerno have the **same** distance (i.e., value= **a** in Table 4) from Naples;
5. Benevento is **less than weakly farther** (i.e., value= **b** in Table 4) than Avellino from Naples;
6. Benevento is **less than weakly farther** (i.e., value= **b** in Table 4) than Salerno from Naples.

The verbal judgements can be synthesized by the following matrix:

$$\mathbf{A^k} = [a_{ij}^k]_{4\times 4} = \begin{bmatrix} & & & \\ \mathbf{c} & & & \\ \mathbf{c} & \mathbf{a} & & \\ \mathbf{c} & \mathbf{b} & \mathbf{b} & \end{bmatrix}.$$

Then, by applying this methodology, we obtain the following eight different matrices (i.e., one for each scale):

Step 1

$$\mathbf{A}^{k1} = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 3 & 1 & 1 & \frac{1}{2} \\ 3 & 1 & 1 & \frac{1}{2} \\ 3 & 2 & 2 & 1 \end{bmatrix} \quad \mathbf{A}^{k4} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 2 & 1 & 1 & \frac{1}{1.58} \\ 2 & 1 & 1 & \frac{1}{1.58} \\ 2 & 1.58 & 1.58 & 1 \end{bmatrix} \quad \mathbf{A}^{k7} = \begin{bmatrix} 1 & \frac{1}{1.5} & \frac{1}{1.5} & \frac{1}{1.29} \\ 1.5 & 1 & 1 & \frac{1}{1.22} \\ 1.5 & 1 & 1 & \frac{1}{1.22} \\ 1.5 & 1.22 & 1.22 & 1 \end{bmatrix}$$

$$\mathbf{A}^{k2} = \begin{bmatrix} 1 & \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ 9 & 1 & 1 & \frac{1}{4} \\ 9 & 1 & 1 & \frac{1}{4} \\ 9 & 4 & 4 & 1 \end{bmatrix} \quad \mathbf{A}^{k5} = \begin{bmatrix} 1 & \frac{1}{1.73} & \frac{1}{1.73} & \frac{1}{1.73} \\ 1.73 & 1 & 1 & \frac{1}{1.41} \\ 1.73 & 1 & 1 & \frac{1}{1.41} \\ 1.73 & 1.41 & 1.41 & 1 \end{bmatrix} \quad \mathbf{A}^{k8} = \begin{bmatrix} 1 & \frac{1}{1.73} & \frac{1}{1.73} & \frac{1}{1.73} \\ 1.73 & 1 & 1 & \frac{1}{1.32} \\ 1.73 & 1 & 1 & \frac{1}{1.32} \\ 1.73 & 1.32 & 1.32 & 1 \end{bmatrix}$$

$$\mathbf{A}^{k3} = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 4 & 1 & 1 & \frac{1}{2} \\ 4 & 1 & 1 & \frac{1}{2} \\ 4 & 2 & 2 & 1 \end{bmatrix} \quad \mathbf{A}^{k6} = \begin{bmatrix} 1 & \frac{1}{1.29} & \frac{1}{1.29} & \frac{1}{1.29} \\ 1.29 & 1 & 1 & \frac{1}{1.13} \\ 1.29 & 1 & 1 & \frac{1}{1.13} \\ 1.29 & 1.13 & 1.13 & 1 \end{bmatrix};$$

Step 2

$$\begin{aligned} &d(\mathbf{A}, \mathbf{A}^{k1}) = 1.5, \quad d(\mathbf{A}, \mathbf{A}^{k2}) = 3.27, \quad d(\mathbf{A}, \mathbf{A}^{k3}) = 1.73, \\ &d(\mathbf{A}, \mathbf{A}^{k4}) = 1.15, \quad d(\mathbf{A}, \mathbf{A}^{k5}) = 1.11, \quad d(\mathbf{A}, \mathbf{A}^{k6}) = 1.24, \\ &d(\mathbf{A}, \mathbf{A}^{k7}) = 1.14, \quad d(\mathbf{A}, \mathbf{A}^{k8}) = 1.09. \end{aligned}$$

Thus, for this student, the less suitable scale (i.e., the worst fitting scale) is the power scale because the matching error $d(\mathbf{A}, \mathbf{A}^{k2}) = 3.27$ is the biggest, and the most suitable scale is the balanced power scale because the matching error $d(\mathbf{A}, \mathbf{A}^{k8}) = 1.09$ is the smallest; note that there is a very small difference among logarithmic, root square, balanced and balanced power scales.

Table 5 Counting the best matching scales for each problem

	P_A	P_B	P_C
1. Linear ($\alpha = 1$)	0	0	0
2. Power ($\alpha = 2$)	0	0	0
3. Geometric ($\alpha = 2$)	0	0	1
4. Logarithmic ($\alpha = 2$)	0	4	3
5. Root square ($\alpha = 2$)	15	18	4
6. Inverse linear	28	28	37
7. Balanced	4	1	4
8. Balanced power ($n = 9$)	5	1	3

Table 6 Averages (geometric means) of the errors obtained in problems P_A, P_B and P_C for each scale

	P_A	P_B	P_C
1. Linear ($\alpha = 1$)	2.135	2.766	2.274
2. Power ($\alpha = 2$)	6.628	13.483	9.760
3. Geometric ($\alpha = 2$)	4.402	11.823	10.435
4. Logarithmic ($\alpha = 2$)	1.371	1.506	2.122
5. Root square ($\alpha = 2$)	1.244	1.401	2.071
6. Inverse linear	1.196	1.406	1.795
7. Balanced	1.314	1.639	1.839
8. Balanced power ($n = 9$)	1.475	1.957	1.915

3.3 Results and discussion

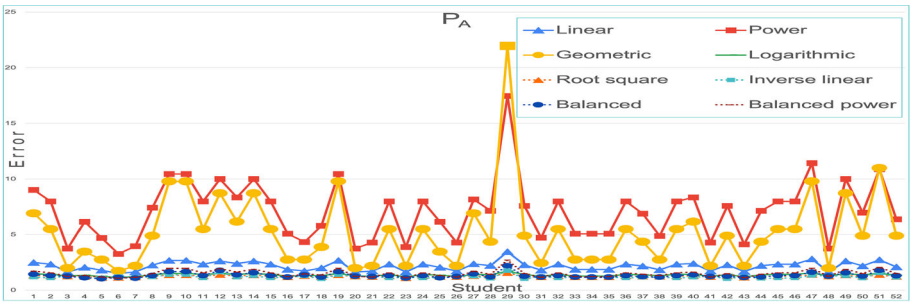
We discarded about 20% of the surveys; that is, 14 surveys where the evaluations were not transitive or an incorrect ranking of the cities was provided. The selection of the wrong city shows incoherence or an error that does not depend on the kind of scale in Table 3. Thus, we analyze $m = 52$ surveys.

By performing *Step 1* and *Step 2* described in the previous section, and by counting how many times a scale provides the smallest matching errors (7), (8) and (9), we obtain Table 5; it shows that the inverse linear scale is always the best fitting scale, followed by the root square scale. As obtained by Meesariganda and Ishizaka (2017) for a strategy selection problem, we obtain that the traditional linear scale used by Saaty and the power and geometric scales provide the worst matching verbal representation of the students.

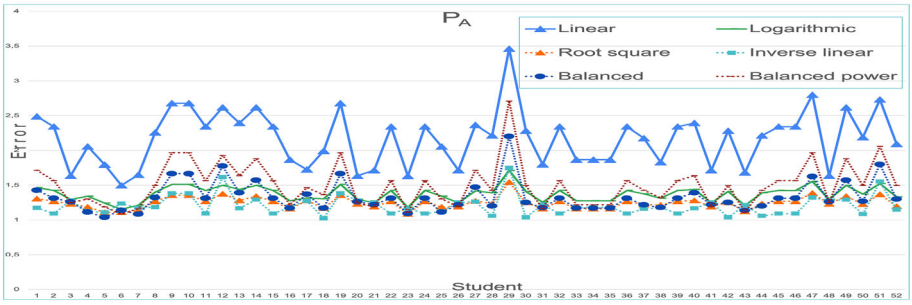
Figures 2, 3, and 4 provide further information, in addition to the count provided in Table 5. In particular, it shows that although the linear scale has never been found to be the best matching verbal representation of the participants, its matching errors are less than the errors provided by the power and geometric scales. In order to highlight the scales with the best matching verbal representations, Figures 2b, 3b and 4b provide a zoom on all the scales with the exclusion of the worst scales (i.e. geometric and power).

In order to analyze the matching errors averages, we provide Table 6; in particular, it confirms that the average error of the linear scale is less than the average errors of the power and geometric scales.

To analyze if the differences among the means in Table 6 are significant, we performed ANalysis Of VAriance (ANOVA) tests. Table 7 shows that the differences among all the scales are significant with an ANOVA test at a confidence threshold $p_t = 0.05$. Indeed, p

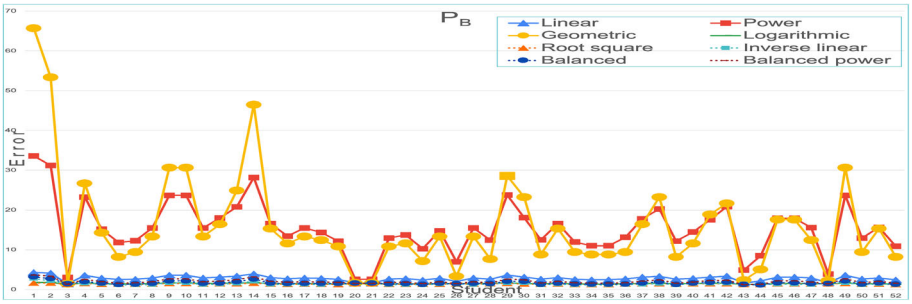


(a) All the scales.

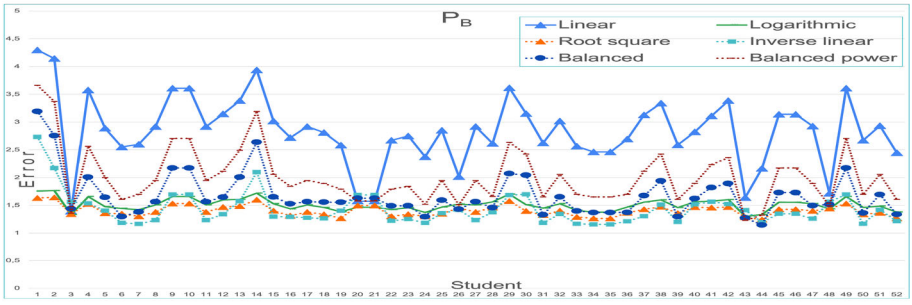


(b) Removing geometric and power scales.

Fig. 2 Matching errors in problem P_A computed by applying (7)

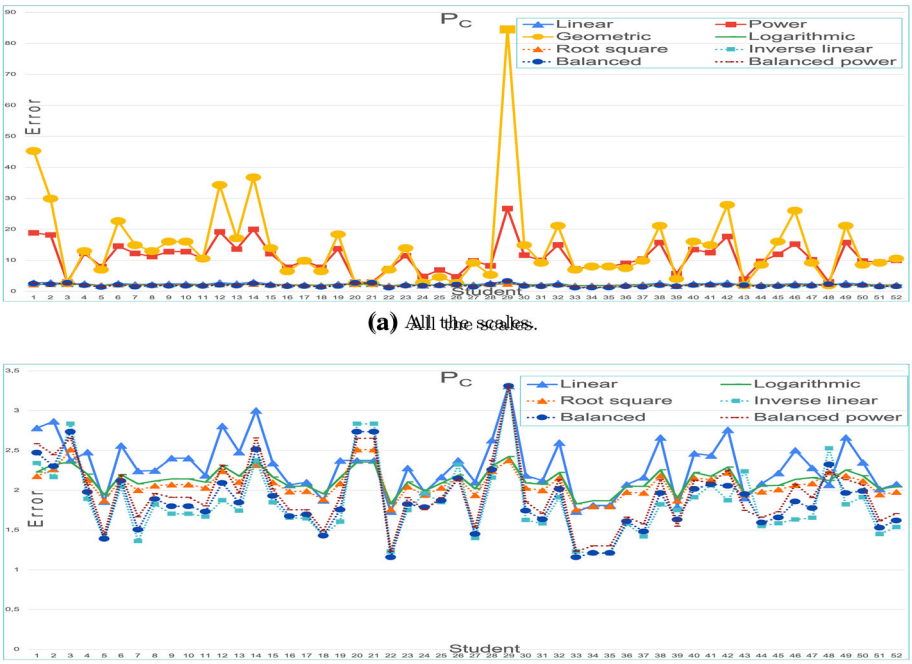


(a) All the scales.



(b) Removing geometric and power scales.

Fig. 3 Matching errors in problem P_B computed by applying (8)



(a) All the scales.

(b) Removing geometric and power scales.

Fig. 4 Matching errors in problem P_C computed by applying (9)

Table 7 ANOVA results

$F_{crit} = 2.032$	P_A	P_B	P_C
p	0	0	0
F	108.368	83.337	49.787

value is always smaller than p_t and F value is always higher than critical value F_{crit}. In particular, there are highly significant differences among all the scales because $p = 0$.

In order to know exactly which means are significantly different from the other ones, we performed a pairwise multiple comparisons; the results are shown in Table 8.

Tables 6 and 8 show that the geometric scale (3.) provides a smaller matching error mean than the power scale (2.) in both P_A and P_B, but not in P_C, and that there is a significant difference between the two means in P_A and no significant differences in P_B and P_C. Note that the difference between these two means and the other ones is always significant. Thus, we can conclude that the power scale provides the worst matching verbal representation of the respondents.

Let us focus now on the scales with the smallest matching errors averages. Tables 6 and 8 show that the inverse linear scale provides the smallest matching errors mean in both P_A and P_C; moreover, in P_A there is not significant difference with the root square scale, and in P_C there is not significant difference with the balanced scale and the balanced power scale. Concerning P_B, the inverse linear and the root square provide the smallest matching errors means and there is not significant difference between them.

Table 8 Pairwise multiple comparisons. Y= there is significant difference. N= there is not significant difference. Numbering of the scales is given in Table 6

	1.	2.	3.	4.	5.	6.	7.
	$P_A P_B P_C$	$P_A P_B P_C$	$P_A P_B P_C$	$P_A P_B P_C$	$P_A P_B P_C$	$P_A P_B P_C$	$P_A P_B P_C$
2.	**Y Y Y**						
3.	**Y Y Y**	**Y** N N					
4.	**Y Y Y**	**Y Y Y**	**Y Y Y**				
5.	**Y Y Y**	**Y Y Y**	**Y Y Y**	**Y Y** N			
6.	**Y Y Y**	**Y Y Y**	**Y Y Y**	**Y** N **Y**	N N **Y**		
7.	**Y Y Y**	**Y Y Y**	**Y Y Y**	N **Y Y**	**Y Y Y**	**Y Y** N	
8.	**Y Y Y**	**Y Y Y**	**Y Y Y**	**Y Y Y**	**Y Y** N	**Y Y** N	**Y Y** N

Thus, we can confirm the results provided in Table 5: the inverse linear scale provides the best matching verbal representation of the respondents.

4 Conclusions

Pairwise comparisons have been used extensively for evaluations in decision-making. However, because they often use linguistic evaluations, conversion to a numerical scale is not obvious. In this paper we use an opinion survey to evaluate the suitability of eight numerical scales to represent the decision maker's verbal representations in problems of different sizes. For this purpose, we compare the scales against known measurable evaluations (i.e. distances between cities). By analyzing the results by means of ANOVA tests, under our methodological choices, we found that the power scale provides the worst matching verbal representation, followed by the geometric scale. The inverse linear scale provides the best matching verbal representation. However, problem-specific scales can be used if the decision-maker has exogenous information on the likely distribution of the pairwise comparisons.

Different parameters on the scales and different distance functions can also be evaluated in a future work. Experiments has been proven to be effective in psychology, economics and recently in behavioral operation research (such as in this paper). In the future, we plan to design more experiments to validate subjective decision techniques; for example, to investigate the suitability of the same numerical scales for subjective decision-making problems.

Funding Open access funding provided by Universitá degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

References

Barzilai, J. (1998). Consistency measures for pairwise comparison matrices. *Journal of Multi-Criteria Decision Analysis, 7*(3), 123–132.

Bozóki, S., Dezsö, L., Poesz, A., et al. (2013). Analysis of pairwise comparison matrices: An empirical research. *Annals of Operations Research, 211*(1), 511–528.

Brunelli, M., & Cavallo, B. (2020). Distance-based measures of incoherence for pairwise comparisons. *Knowledge-Based Systems, 187*, 104808.

Brunelli, M., & Cavallo, B. (2020). Incoherence measures and relations between coherence conditions for pairwise comparisons. *Decisions in Economics and Finance, 43*(2), 613–635.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*(3), 391–405.

Cavallo, B. (2019). $\mathcal{G}$-distance and $\mathcal{G}$-decomposition for improving $\mathcal{G}$-consistency of a pairwise comparison matrix. *Fuzzy Optimization and Decision Making, 18*(1), 57–83.

Cavallo, B., & D'Apuzzo, L. (2009). A general unified framework for pairwise comparison matrices in multicriterial methods. *International Journal of Intelligent Systems, 24*(4), 377–398.

Cavallo, B., & D'Apuzzo, L. (2020). Relations between coherence conditions and row orders in pairwise comparison matrices. *Decisions in Economics and Finance, 43*(2), 637–656.

Cavallo, B., Ishizaka, A., Olivieri, M. G., et al. (2019). Comparing inconsistency of pairwise comparison matrices depending on entries. *Journal of the Operational Research Society, 70*(5), 842–850.

Csató, L. (2017). On the ranking of a Swiss system chess team tournament. *Annals of Operations Research, 254*(1–2), 17–36.

Csató, L. (2018). Characterization of an inconsistency ranking for pairwise comparison matrices. *Annals of Operations Research, 261*(1–2), 155–165.

Csató, L. (2019). Axiomatizations of inconsistency indices for triads. *Annals of Operations Research, 280*(1–2), 99–110.

Donegan, H. A., Dodd, F. J., & McMaster, T. B. M. (1992). A new approach to AHP decision-making. *Journal of the Royal Statistical Society: Series D (The Statistician), 41*(3), 295–302.

Dong, Y., Hong, W. C., Xu, Y., et al. (2013). Numerical scales generated individually for analytic hierarchy process. *European Journal of Operational Research, 229*(3), 654–662.

Donohue, K., Katok, E., & Leider, S. (2018). *The handbook of behavioral operations*. Hoboken: Wiley.

Elliott, M. A. (2010). Selecting numerical scales for pairwise comparisons. *Reliability Engineering & System Safety, 95*(7), 750–763.

Harker, P. T., & Vargas, L. G. (1987). The theory of ratio scale estimation: Saaty's analytic hierarchy process. *Management Science, 33*(11), 1383–1403.

Huizingh, E. K., & Vrolijk, H. C. (1997). A comparison of verbal and numerical judgments in the Analytic Hierarchy Process. *Organizational Behavior and Human Decision Processes, 70*(3), 237–247.

Ishizaka, A., & Labib, A. (2009). Analytic hierarchy process and expert choice: Benefits and limitations. *OR Insight, 22*(4), 201–220.

Ishizaka, A., & Nguyen, N. H. (2013). Calibrated fuzzy AHP for current bank account selection. *Expert Systems with Applications, 40*(9), 3775–3783.

Ishizaka, A., & Siraj, S. (2018). Are multi-criteria decision-making tools useful? An experimental comparative study of three methods. *European Journal of Operational Research, 264*(2), 462–471.

Ishizaka, A., & Siraj, S. (2020). Interactive consistency correction in the analytic hierarchy process to preserve ranks. *Decisions in Economics and Finance, 43*(2), 443–464.

Ishizaka, A., Balkenborg, D., & Kaplan, T. (2011). Influence of aggregation and measurement scale on ranking a compromise alternative in AHP. *Journal of the Operational Research Society, 62*(4), 700–710.

Kagel, J., & Roth, A. (2017). *The handbook of experimental economics* (Vol. 2). Princeton: Princeton University Press.

Keeney, R. L., von Winterfeldt, D., & Eppel, T. (1990). Eliciting public values for complex policy decisions. *Management Science, 36*(9), 1011–1030.

Liang, L., Wang, G., Hua, Z., et al. (2008). Mapping verbal responses to numerical scales in the analytic hierarchy process. *Socio-Economic Planning Sciences, 42*(1), 46–55.

Linares, P. (2009). Are inconsistent decisions better? An experiment with pairwise comparisons. *European Journal of Operational Research, 193*(2), 492–498.

Lootsma, F. (1989). Conflict resolution via pairwise comparison of concessions. *European Journal of Operational Research, 40*(1), 109–116.

Ma D, Zheng X (1991) 9/9-9/1 scale method of AHP. In: *2nd Int. Symposium on AHP*, pp 197–202.

Meesariganda, B. R., & Ishizaka, A. (2017). Mapping verbal AHP scale to numerical scale for cloud computing strategy selection. *Applied Soft Computing, 53*, 111–118.

Millet, I. (1997). The effectiveness of alternative preference elicitation methods in the analytic hierarchy process. *Journal of Multi-Criteria Decision Analysis, 6*(1), 41–51.
Por, H. H., & Budescu, D. V. (2017). Eliciting subjective probabilities through pair-wise comparisons. *Journal of Behavioral Decision Making, 30*(2), 181–196.
Ramík, J. (2015). Isomorphisms between fuzzy pairwise comparison matrices. *Fuzzy Optimization and Decision Making, 14*(2), 199–209.
Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega, 53*, 49–57.
Rokou, E., & Kirytopoulos, K. (2014). Calibrated group decision process. *Group Decision and Negotiation, 23*, 1369–1384.
Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 15*(3), 234–281.
Salo, A. A., & Hämäläinen, R. P. (1997). On the measurement of preferences in the analytic hierarchy process. *Journal of Multi-Criteria Decision Analysis, 6*(6), 309–319.
Tanino, T. (1984). Fuzzy preference orderings in group decision making. *Fuzzy Sets and Systems, 12*(2), 117–131.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.
Whitaker, R. (2007). Validation examples of the analytic hierarchy process and analytic network process. *Mathematical and Computer Modelling, 46*(7), 840–859.
Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364.
Wixted, J. (2018). *Stevens' handbook of experimental psychology and cognitive neuroscience*. Hoboken: Wiley.
Yokoyama, M. (1921). The nature of the affective judgment in the method of paired comparisons. *The American Journal of Psychology, 32*(3), 357–369.