Methodological Aspects of Measuring Preferences Using the Rank and Thurstone Scale

Joanna Dębicka | *Wroclaw University of Economics and Business, Wroclaw, Poland*
Edyta Mazurek[1] | *Wroclaw University of Economics and Business, Wroclaw, Poland*
Katarzyna Ostasiewicz | *Wroclaw University of Economics and Business, Wroclaw, Poland*

Received 31.1.2022 (revision received 1.4.2022), **Accepted** (reviewed) 4.4.2022, **Published** 16.9.2022

Abstract

The fundamental problem with the measurement of preferences is that it not only attempts to measure something that is, by its nature, "unmeasurable", but also hidden from a direct observation. In addition, a person's current emotional, material and social situation influences the measurement of preferences resulting from the person's system of values. The paper is a study on the methodology of preference measurement, a comparison and evaluation of two methods of scale construction. Among various techniques we investigate the two methods: Thurstone procedure for finding scale separations developed by Thurstone and the simplest rank method of scaling. This study examines the relative merits of Thurstone and rank techniques of scale construction.

Keywords	***DOI***	***JEL code***
Preference ranking, Case V, Thurstone model, structure similarity test, input matrix	*https://doi.org/10.54694/stat.2022.5*	*C18, C83, C46*

INTRODUCTION

Questionnaire survey research became very popular in scientific research. In social and psychological research it is often used to describe and explore human behavior (Singleton and Straits, 2018). Moreover, this kind of learning about social preferences is frequently used by policymakers. The presidential ballot is a kind of "survey", in which each citizen is asked about his preferences for the person to be the head of the country. In more common situations citizens are asked to vote for the projects to be addressed by civic funds. While willing to eliminate the barriers that keep disabled persons from full participation in social and civic life, it is worth knowing, which barriers are most burdensome for most of them.

The shift from authoritative decision making to public consultations has been included in the EU's research and innovation program *Horizon Europe 2021–2027* (*Horizon Europe, European Commission*).

[1] Wroclaw University of Economics and Business, Department of Statistics, Komandorska 118/120, Wroclaw, Poland. Corresponding author: e-mail: edyta.mazurek@ue.wroc.pl.

This research agenda has been defined in terms of five missions to be carried out. One of those missions concerns is the adaptation to social transformation consisting directly to involving all actors, including citizens, civil society organizations, and public authorities, in research, innovation and change. One such change/transformation concerns the reduction of socio-economic inequalities. Co-creation is the preferred method of analyzing changes (as well as a key aspect of research and implementation of solutions) in *Horizon Europe 2021–2027*. In this context, the issues of identifying the most necessary changes (ranking individual preferences and building scales) and defining specific actions for priority areas of change (questionnaire research on specific solutions, including the impact of the method of asking a question (type of question) on the answer) become important.

Although most of the researchers agree on the importance of survey research, there are many doubts and controversies concerning the methods of asking questions and, after collecting data, of aggregating answers into something that might be regarded as the collective preferences and choices (Holbrook, Cho and Johnson, 2006). The problem has two most important components. The first is a mainly psychological matter of how to ask questions to make people reveal their true attitudes. Still, some psychologists claim, that surveyed persons do not reveal but rather construct their attitudes in the process of being surveyed. It is known, that websites and online survey software are on the one hand useful to assist in the design and delivery of questionnaires, but, on the other hand, they can also introduce sources of bias (Ball, 2019).

The second crucial issue is the aggregation of preferences, and this is the particular branch of survey studies to which our paper is contributing. It focuses on the methodology of aggregated preference measurement, a comparison and evaluation of two methods of scale construction. Among various techniques, we investigate the two methods: Thurstone procedure for finding scale separations developed by Thurstone and the simplest rank method of scaling.

Stepping back to XVIII century for the discussion between Borda and Condorcet about the best method of aggregating preferences in voting systems, which by now has not found the conclusive solution, one may say that the issue, which method of aggregation is the very best, is a vague and to some degree unscientific but also an axiological question. Still, it is worth comparing different kinds of aggregation methods, to be conscious of potential differences and characteristics. For example, it is worth knowing if different methods give qualitatively different results. If so, it is of crucial importance to consider very thoroughly which method is better for a given aim and why. If the results are similar, it might be of use to investigate more technical properties – stability of the results, sensitivity for individual observations and so on.

Thurstone scaling is the well-known tool for the estimation of preferences among objects by the observed frequencies of their paired comparisons (Thurstone, 1927a; Thurstone and Chave, 1929; Thurstone and Jones, 1957). The positioning of items on this scale can be found by averaging the percentiles of the standard normal distribution corresponding to the proportions of the respondents preferring one item over each of the others. This scaling is widely used in applied psychology, particularly in marketing and advertising research (Edwards and Kenney, 1946; Escher, 2010). Statistical approaches to the Thurstone scaling were considered by Mosteller (Mosteller, 1951), and various modifications of this model were developed by Lipovetsky (Lipovetsky, 2007), and Saffir (Saffir, 1937). The authors made comparison of the methods of attitude scale construction of Thurstone, Likert, and Guttman and Bradley-Terry model (Edwards and Kenney, 1946; Edwards and Kilpatrick, 1948; Lipovetsky and Conklin, 2004; Drasgow, Chernyshenko and Stark, 2010; Tsukida and Gupta, 2011; Stadthagen-González et al., 2018; Edwards and Kilpatrick, 1948). When the number of stimuli is large, the number of pairs to be compared becomes very large, and the similarity task is inefficient. Tsogo, Masson and Bardot reviewed the main similarity task methods suitable for large sets of objects (Tsogo, Masson and Bardot, 2000). They point out the advantages and disadvantages of such methods as: incomplete similarity tasks, binary dissimilarities, hierarchical sorting tasks, conditional rank-order. Among the comparisons, there was no comparison with the classical approach based on the sum of ranks. We decided to compare Thurstone scale and direct rank method of scaling. In the paper, the words project, object or stimulus, are used interchangeably and symbolically denote a ranking object.

The paper is organized as follows. In Section 1 we give information about the classification of scaling techniques and describe the rank scale and Thurston method in detail. Section 2 consists of results obtained from the survey, which was constructed based on the original Thurstone study that measured social values, specifically the seriousness of different types of crimes or offences. In Section 2.1. we describe the survey, dataset and give rankings of offences, from the worst to the lightest. Section 2.2. offers a comparative analysis of the crime severity scales obtained by ranking methods and the Thurstone scale. In Section 2.3 we check the assumption of the independence of alternatives for subsets of crime and offences. Finally, last Section sums up results obtained in the paper indicating also the direction of future research.

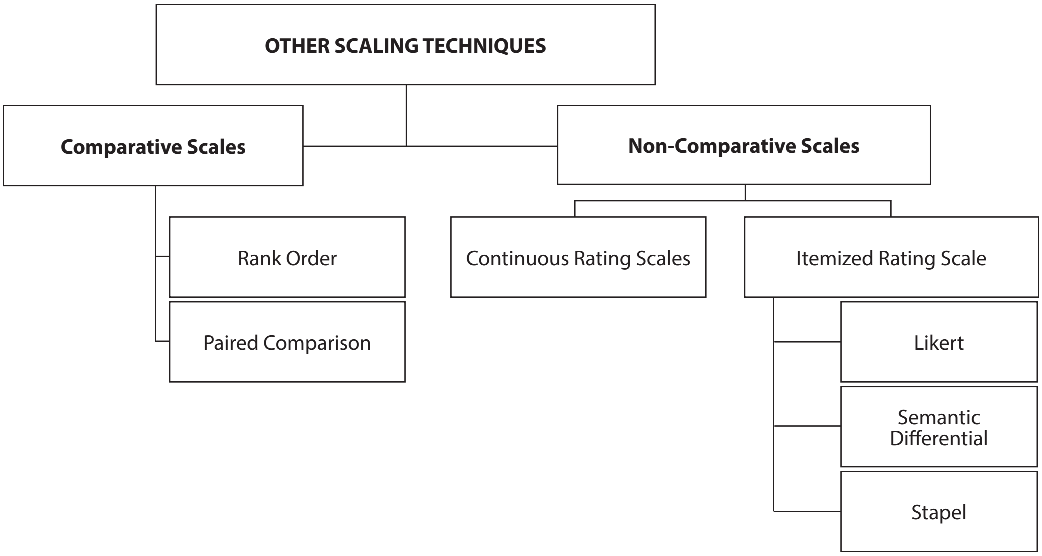
1 SCALING TECHNIQUES

Scaling objects can be used for a comparative study of more than one object. For a long time marketing and some other kinds of researches have been highly dependent on those techniques. Scaling emerged from the social sciences in an attempt to order attributes with respect to quantitative attributes. Scaling provides a mechanism for measuring abstract concepts.

1.1 Classification

In general, scaling techniques can be divided into two categories: non-comparative scales and comparative scales (see Figure 1). A non-comparative scale is used to analyze an individual product or object's performance on different parameters and is most frequently used in marketing research. In this approach each object is scaled independently on the others, e.g., respondents may be asked how they are satisfied with product A, product B, etc., without comparing the product. Contrary, within comparative scales, respondents are asked to place one object regarding other objects.

Figure 1 Classification of scaling techniques



Source: Own elaboration, following standard textbook presentation

Our research focuses on ranking, so on ordinal tasks. Ordinal tasks involve ranking objects in some way to produce dominance data; that is, one stimulus dominates another, so only judgments of greater than or less than are required. Ranking can be accomplished directly or derived from pairing the objects. A paired comparison symbolizes two objects from which the respondent needs to select one according to their preference. The direct ranking consists of assigning integers to objects, indicating the order of preferences. These two methods of creating ranking are most commonly used in practice. For that reason, the paper refers to comparing two methods of attitude scale construction. Both rank scale and Thurstone method belong to comparative scales. The difference in classification is, that the rank scale can be accomplished only by rank order, while Thurstone method can be applied in the case of both pairwise comparisons and by rank order (from which pairwise comparisons can be obtained).

1.2 Rank scale

The method of determining the scale based on the assigned ranks will be presented in an example (see Table 1). Let us assume that we have five respondents (so-called judges). Each respondent orders crimes A-E from the worst to the lightest. Then for each offence, we set the sum of ranks. In the end, we re-scale it to the range 0–1 using min-max normalization. According to this example, crime E is the worst and crime A is the lightest. Eventually, we assigned a rank and a value from the range [0.1] for each crime.

Table 1 Example of the scale based on the assigned ranks

Judge	Offences				
	A	B	C	D	E
1	5	3	4	2	1
2	3	4	2	5	1
3	4	5	3	2	1
4	5	3	4	1	2
5	5	2	3	4	1
Sum of ranks	22	17	16	14	6
Min-max scaling	1	0.69	0.63	0.5	0

Source: Own elaboration

The disadvantage of this approach is decreasing efficiency while increasing the number of evaluated objects. Moreover, we assume that the distances between objects, considered as the validity of one object over the second one, is equal. On the other hand, the great advantage of this approach is its simplicity and lack of assumptions.

1.3 Thurstone scaling (Case V)

Thurstone pair comparison model is considered a probabilistic choices model with the following assumptions:

1. Distribution of the hidden preferences in the preferences had a normal distribution.
2. Preferences are independent of each other, and they have one source of variance (the assumption that there is zero correlation might be softened to the assumption that there is a correlation between pairs).
3. The probability of the intransitive preferences is different from zero.
4. Measurements errors are non-correlated, and they have a normal distribution.

Thurstone (1927b) assumed and provided a rationale for ordering objects on a continuum. Although we may have more or less favourable reactions to a particular object, Thurstone suggested that there was a most frequent or typical reaction to any object. Because the normal curve is symmetrical, the most frequent reaction occupies the same scale position as the mean. Thus, the mean can also represent the scale value for the particular object. So, in his simplest Thurstone model (so-called Case V), he assumed that reactions to various stimuli were normally distributed. He also assumed that the variance of the reactions around each mean would be the same. The means of normal distribution of each object are interpreted as scale values.

In the method of pairs comparison: for n objects, we get $\frac{n(n-1)}{2}$ pairs. Let X_i $(i = 1, 2, \ldots, n)$ be the characteristic of an object. We assume that $X_i \sim N(\mu_i, \sigma_i)$.

Parameter μ_i is an expected value of the ith object and is the main topic of interest in the current context, as we want to compare the relative positions of the objects, i.e. their central tendencies. Estimation μ_i, as an item on a scale, is based on observation of the difference $X_i - X_j$. Note that a random variable $Y = Y_{ij} = X_i - X_j$ has a normal distribution with the following density function:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp\left(-\frac{(y-(\mu_i-\mu_j))^2}{\sqrt{2\pi\sigma_{ij}^2}}\right), \tag{1}$$

where: $\sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_i\sigma_j\rho_{ij}}$.

Through the comparison of projects in pairs, (X_i, X_j)it is possible to determine the probability estimator:

$$p_{ij} = P(X_i - X_j > 0) = \Phi\left(\frac{\mu_i - \mu_j}{\sigma_{ij}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\mu_i-\mu_j)/\sigma_{ij}} e^{-y^2/2} dy\,. \tag{2}$$

Knowing p_{ij} it is possible to determine $z_{ij} = \Phi^{-1}(p_{ij})$. Then by using the least-squares approximation $(\sum_{i \neq j}^{m} (z_{ij} - (\mu_i - \mu_j))^2 \rightarrow \min)$ we determine $\mu_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij} = \frac{1}{n} \sum_{i=1}^{n} \Phi^{-1}(p_{ij})$.

After re-scaling μ_j to the range 0-1 using min-max normalization:

$$\mu_j' = \frac{\mu_j - \min_j(\mu_j)}{\max_j(\mu_j) - \min_j(\mu_j)}, \tag{3}$$

the average values μ create a Thurston preferences scale that is most commonly scale for range [0,1].

As it could be seen Thurstone scale is based on some particular assumptions. It seems legitimate to inquire whether it works better than the simpler scale that may be used and whether it is possible to construct equally reliable scales without making unnecessary statistical assumptions. To this aim in the next section simple rank scale will be introduced.

2 SURVEY RESEARCH – EXPERIMENT

2.1 Description survey, dataset and scales

The survey was constructed based on the original Thurstone study that measured social values, specifically the seriousness of different types of crimes or offences. The following types of crimes/offenses were considered in the survey:

P1. Violent rape.
P2. Assault with a severe body injury.
P3. Paedophile acts.
P4. Domestic violence.
P5. Threats.
P6. Murder.
P7. Defamation (slander).
P8. Harassment.
P9. Kidnapping for ransom.
P10. Identity theft.

We shall assume the seriousness of an offence to be the seriousness as judged rather than as measured in terms of objective consequences or in some normative way. The main aim of the study was to obtain data for comparative analysis. The intermediate aim was to perform some kind of pilot study as a preliminary step to learn about societal preferences regarding the strength of the crime. In Poland, public dissatisfaction resulting from inadequate punishment for crime is often heard in discussions. It could be useful to have some knowledge of societal judgments of sentences for given crimes. The pilot study could serve as a useful tool to project the actual survey, especially the final set of crimes to be included. Of course, the final survey would have to be carried out on a representative sample of the population.

The respondents were 219 students (individuals aged 19–23) in the conducted study. Students responded in two ways. The first way was that the offences were arranged in pairs so that they were paired with every other one. The total number of pairs of offences presented was 10(10 – 1)/2 = 45. A student had to choose a more severe crime from each pair. This method excludes the draw situation. Hence, if a student considered crimes equally serious, they have to choose one of them as worse. The input matrix **P** (matrix observed proportion of times that object *i* was chosen over object *j*) obtained based on the data is presented in the form of Table 2, where e.g. $p_{21} = P(P2 \succ P1)$ means that 18% of respondents considered that the P2 offense (assault with a serious body injury) is more serious than the P1 offence (violent rape) and $p_{12} = P(P1 \succ P2)$ means that 82% of respondents answered the opposite, that the P1

Table 2 The input matrix and Thurstone scale for survey data

P	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	0	0.82	0.33	0.82	0.99	0.33	0.99	0.97	0.94	0.92
P2	0.18	0	0.31	0.64	1.00	0.13	0.99	0.96	0.86	0.89
P3	0.67	0.69	0	0.84	0.98	0.37	0.98	0.98	0.83	0.93
P4	0.18	0.36	0.16	0	0.96	0.11	0.95	0.92	0.50	0.80
P5	0.01	0.00	0.02	0.04	0	0.02	0.60	0.27	0.05	0.21
P6	0.67	0.87	0.63	0.89	0.98	0	1.00	0.97	0.95	0.95
P7	0.01	0.01	0.02	0.05	0.40	0.00	0	0.22	0.06	0.19
P8	0.03	0.04	0.02	0.08	0.73	0.03	0.78	0	0.11	0.42
P9	0.06	0.14	0.17	0.50	0.95	0.05	0.94	0.89	0	0.77
P10	0.08	0.11	0.07	0.20	0.79	0.05	0.81	0.58	0.23	0
					↓					
Thurstone scaling	*0.09*	*0.23*	*0.10*	*0.40*	*0.98*	*0*	*1*	*0.81*	*0.48*	*0.68*

Source: Own elaboration

crime is more severe than the P2 crime. At the end of Table 2 is the *Thurston scale* (developed as intended and the technique presented in Section 1.2).

Another way of scaling crimes was based on students' ranking of offences, from the worst to the lightest. In that case, the draw situation was not possible, either. The results are given in Table 3.

Table 3 Rankings for survey data

Rank	**Offences**									
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
1	39	0	31	1	0	146	0	0	0	2
2	102	27	58	4	0	21	1	0	3	3
3	61	43	59	13	0	34	1	0	6	2
4	13	76	36	47	1	10	0	3	30	3
5	2	45	21	67	3	5	1	3	61	11
6	1	19	12	59	9	0	6	13	73	27
7	0	7	1	21	23	0	10	57	26	74
8	1	2	0	3	59	0	27	79	11	37
9	0	0	1	3	75	2	40	51	4	43
10	0	0	0	1	49	1	133	13	5	17
Sum of ranks	*502*	*891*	*661*	*1 133*	*1 871*	*383*	*2 022*	*1 725*	*1 243*	*1 614*
Rank	*0.1*	*0.3*	*0.2*	*0.4*	*0.9*	*0*	*1*	*0.7*	*0.5*	*0.6*
Rank scale *Min-max scaling*	*0.07*	*0.31*	*0.17*	*0.46*	*0.91*	*0*	*1*	*0.82*	*0.52*	*0.75*

Source: Own elaboration

The penultimate row of Table 3 contains a scale, or rather no scale, called a *rank*. It shows the situation where we rank the crimes and do not perform the scaling. In that situation, the distances, considered differences between subsequent offences' validity, are the same. The last row of Table 3 contains the ranking combined with the re-scaling the sum of ranks to the range 0–1 using min-max normalization, so-called *rank scale (*cf. Section 1.1).

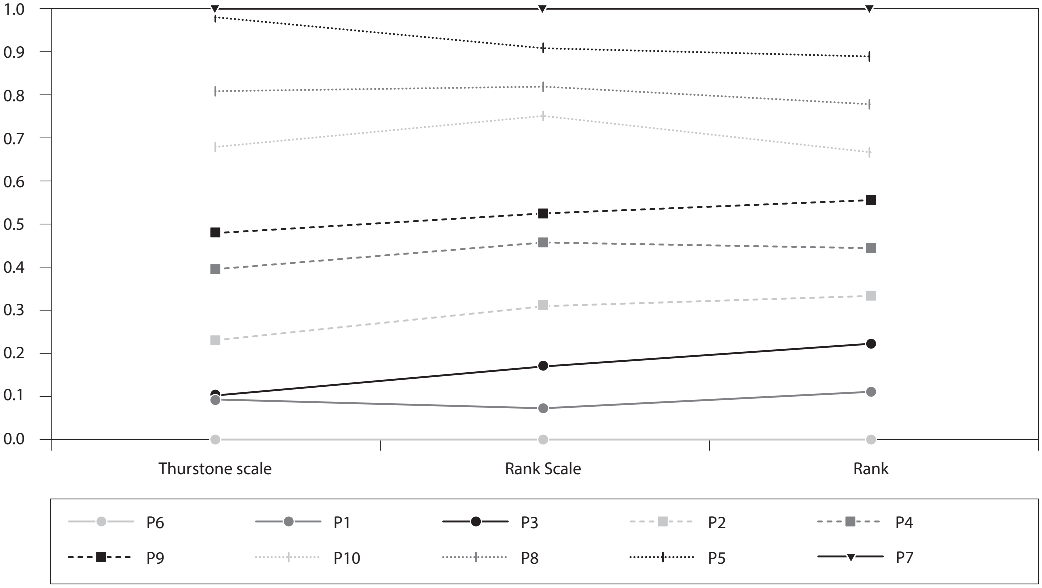
Moreover, students were asked which way of crime assessment was easier and more comfortable for them: ranking or pair comparison. 67% of responders said it was easier to rank the crimes instead of pair comparison. However, 33% of them preferred to evaluate pairs. It can be assumed that the predominance of respondents preferring to rank crimes will increase with the increase in the number of crimes (or objects to compare).

2.2 Comparative analysis of scales

First, let's compare the crimes' scale obtained by the ranking methods and the Thurstone scale. Hence, we have to ask ourselves: *How do the offences arrange themselves in a quantitative continuum from those that seem to be most serious to those that seem relatively least objectionable?* Figure 2 is a graphic illustration of the Thurston scale and rank scale, and the simplest ranking is also included for comparison.

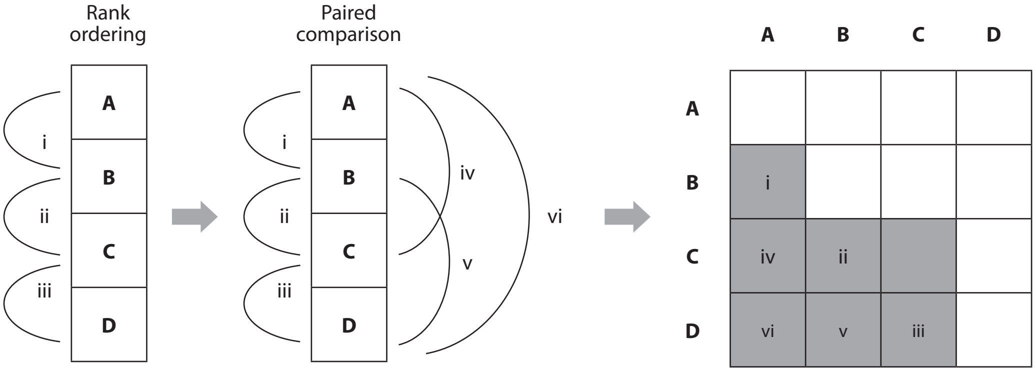
As we can see in the graph in Figure 2, each method shows the same hierarchy of crimes. Here, the results are consistent. In contrast, the differences could be seen in the Thurstone and rank scale results. The greatest difference refers to the crimes: P1 and P3 and P5 and P7. According to the Thurstone scale, crimes in both pairs are equally important. While, on the contrary, the rank scale differs the importance of crimes in these pairs. It means that the choice of method affects the final results. The question arises: *Are the differences statistically significant, and whether the way respondents are asked about the importance of crimes affects the scaling results?*

Figure 2 Comparison of rankings



Source: Own elaboration

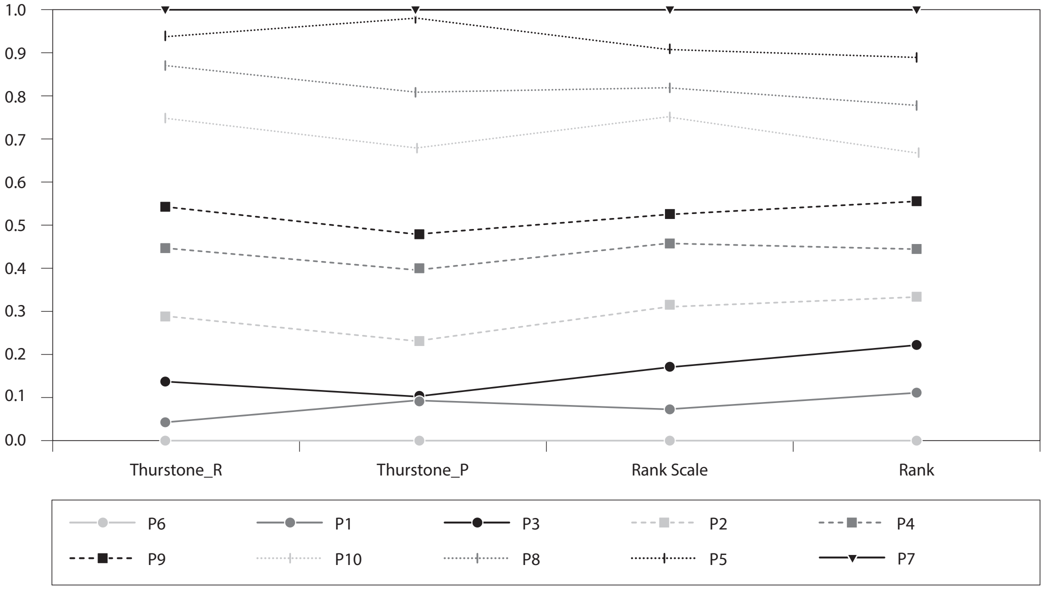
Figure 3 Schematic representation of deriving paired comparison data based on rank data



Source: Own elaboration

The input data for the Thurstone scale is the matrix P, where symmetric cells sum to unity. We could obtain matrix P from the rank order as well as from the paired comparison (see Figure 3). Because in the experiment, the respondents ranked crimes using both methods, so it is possible to compare the results of the Thurstone scale obtained from the rank ordering and paired comparison.

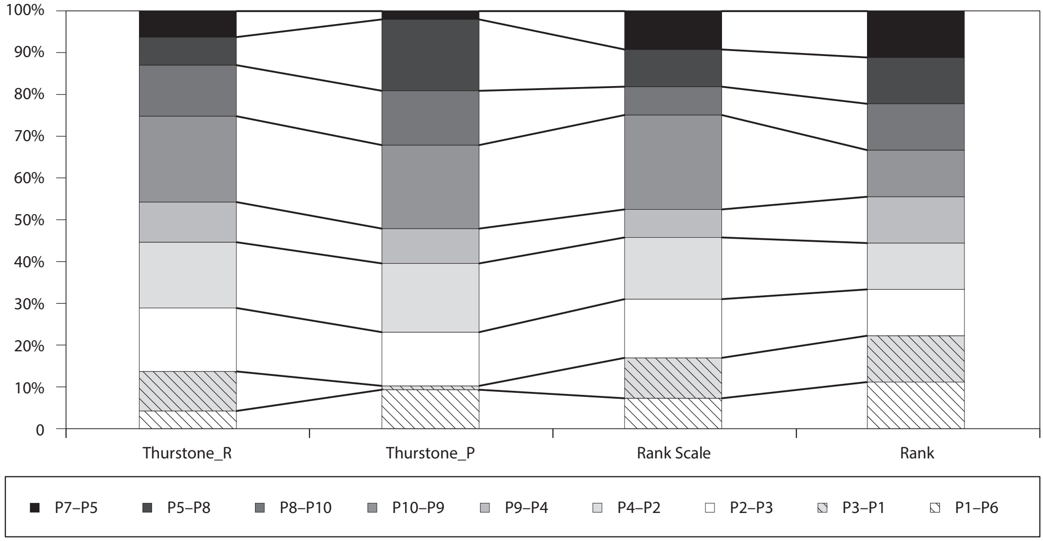
Figure 4 Schematic representation of deriving paired comparison data based on rank data



Source: Own elaboration

Figure 4 shows the influence of the used way of asking about preferences on the scale values. Thurstone_P denotes the Thurstone scale obtained by the paired comparison. Whereas Thurstone_R denotes the Thurstone scale obtained from the rank ordering. For comparison, the classification of crimes has

Figure 5 The structures of the differences between the values for scales



Source: Own elaboration

been included, without their scaling, i.e. rank. Again, the difference between scales could be seen. The scale Thurstone_R, similarly to the rank scale, do not state pairs of crimes: P1, P3 and P5, P7 as of equal importance. It means that the way of asking the respondents about the projects' hierarchy has an impact on the determined scale. However, whether the differences between the obtained results are statistically significant is still valid. We used the test for Similarity of Structures proposed by (Sokołowski, 1993) to answer this question. There are two test hypotheses:

H_0: The structures are dissimilar,

H_1: The structures are similar.

The test statistic is the structural similarity coefficient based on the Bray-Curtis distance, and for a given level of significance, we have a right-tailed rejection region. For each considered scale, the structure was made by the differences between the values on a scale (as in Figure 5).

Table 4 The test similarity of structures of the differences between the values for scales

Compared structures for scales	The value of the test statistic	Significance level	Critical value
Thurstone_P : Thurstone_R	0.83	0.01	0.7705
Thurstone_P : Rank Scale	0.90	0.05	0.7115
Thurstone_R : Rank Scale	0.80	0.10	0.6747
Thurstone_R : Rank	0.81	0.16	0.80
Thurstone_P : Rank	0.76	0.02	0.21
Rank Scale : Rank	0.82	0.63	0.95

Source: Own elaboration

For almost all pairs of the structures compared, the test rejected the null hypothesis, which means that each method showed a similar scale of crimes (cf. Table 4). The exception is comparing Thurstone_P and the rank on significance level equal 0.1. Whereas the test result, it turned out that the distance structures between successive distances of the Thurstone_P scale differ significantly from those obtained without any scaling (rank).

That result shows another direction for further research. Searching for conditions concerning matrix P leading to equality of the Thurstone scale and the rank scale. Moreover, the conclusion can be stated that the differences between obtained results for scaled orders (Thurston_P, Thurston_R and the rank scale) in the set of crimes and offences are not statistically significant.

2.3 The independence of alternatives

Finally, we consider *the problem of the independence on alternatives.* Independence on alternatives means that if P1 is preferred to P2 out of the choice set {P1, P2}, then introducing a third alternative P (thus expanding the choice set to {P1, P2, P}) should not make P2 preferred to P1. Hence, the independence of alternatives assumes that ordering a given pair of items does not depend on the other options available. Changes in individuals' rankings of irrelevant alternatives (ones outside a certain subset) should have no impact on the societal ranking of the subset.

As both the ranking scale and Thurstone method are special kinds of aggregation of preferences, it is known – as proven by Arrow’s Impossibility Theorem – that dependence on irrelevant alternatives cannot be avoided in a general case. Still, we may follow the attempts of researchers who investigate the frequency of occurring the Condorcet paradox in the real data (although, again, this paradox cannot be avoided in principle), and to check the frequency of occurring the dependence on irrelevant alternatives in scaling tasks. Not much effort has been devoted by now to this aim, so our investigation is a contribution and a trigger to such discussion. The very seminal paper of Thurstone (1929) has not discussed that topic, although investigating the data available in this paper one may note the existence of this undesirable property of the data.

We have checked the dependence on alternatives for our data for rank method. Namely, we have investigated the relative positions of all combinations of subsets (starting from 2 ending with 9 elements) to check if their relative positions may change depending on the other objects in the subset under consideration. We conclude, that there is not a single interchange of the relative position of any couple of objects.

Contrary, the graph in Figure 6 shows the counterexamples for Thurstone scale for the three examples of crimes' subsets consisting of:

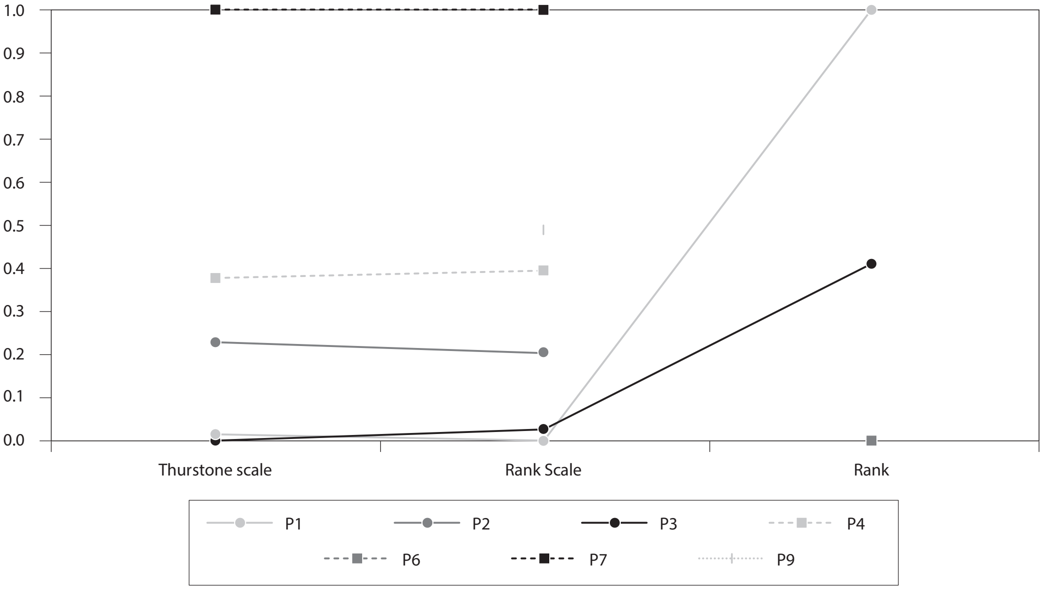
Case 1: {P1, P2, P3, P4, P7},

Case 2: {P1, P2, P3, P4, P7, P9},

Case 3: {P1, P3, P6}.

Note that the second case extends the first case by the crime P9. The third case consists of 2 questions from the previous two cases and additionally P6. The importance of crimes P1 and P2 is influenced by the other crimes considered in the comparison. The inclusion of the offence P9 made the offence P3 lighter than the crime P1.

Figure 6 Thurstone scales for the subgroups of crimes



Source: Own elaboration

We do not infer anything conclusive with this demonstration but rather pose a question. Does it seem sensible to reject those results of finding the aggregated scales in which the dependence on irrelevant alternatives is observed (or at least to reject those objects, which relative positions are sensitive for alternatives)? The lack of this effect of course does not ensure that adding yet another alternative that has not been included in the survey will not change the situation and one is not able to protect from this possibility. Still, the lack of the undesirable effect – while it is not a proof – is at least some corroboration of stability of the results, given the method used to obtain them. Again relating to the Condorcet Paradox – some authors suggest that if the data reveals this undesirable property for one method of revealing the winner, one should switch into another method. Thus, in the case when the data reveal sensitivity for irrelevant alternatives, should one switch to a more robust (but more coarse) method or is it enough to reject the troublesome object? But what in the case when they are crucial for the task at hand and shouldn't be removed without the loss of the usefulness of the results?

While not daring to suggest the definite answer, we think, that investigating the existence and frequency of the occurring of undesirable properties of Thurstone method as compared with more simple ones is a valuable deepening the understanding of this scaling method.

CONCLUSIONS

Among various scaling techniques, the simplest one is probably the rank scaling. The simplicity may be however regarded as oversimplicity. On one hand, an individual is asked to rank objects in an ordinal way, but as a final result, by averaging ranks, we obtain an interval scale, which seems to be a kind of inconsistency. One may argue that this way of obtaining the scale is supported by the assumption of normality (or at least symmetry) of distribution of ranks – that is, the same fraction of respondents ranking object as the second would place it in the vicinity of the first one (rather than in the middle between the first and the third) as the fraction of individuals who would place it in the vicinity of the third object rather than in the equal distance from the first and the third. Thurstone scale is based on the explicit and detailed model of objects, which perception is distributed normal. The estimation of the expected value of each distribution is intermediated by the relative positions of all couples of two objects. Although both methods are prone to some disadvantages – in this paper we have examined the dependence on irrelevant alternatives – it seems that Thurstone method, as a more sophisticated one, is also more susceptible to such effects, at least in the particular case of our empirical study.

We have shown that both methods gave qualitatively similar results in the statistical sense. However, as for precise results, it is still not obvious which scale should be treated as appropriate. One question is the justifiability of the assumption underlying the Thurstone model. The other is the undesirable property of this particular results – dependence on the irrelevant alternatives. It seems that if we want to adopt Thurstone scale as a valid one, we should be conscious of the problem with "unstable" results (specifically, objects P1 and P3), and either remove them from the analysis or treat them as undistinguishable within the given method.

Our study does not propose any definite answer to the question of which method to be used but rather identifies the problem with the supposed unreliability of some results in the case of dependence on irrelevant alternatives – the problem that was bypassed in silence by both Thurstone and the followers.

References

BALL, H. L. (2019). Conducting Online Surveys [online]. *Journal of Human Lactation*, 35(3): 413–417. <https://doi.org/10.1177/0890334419848734>.

DRASGOW, F., CHERNYSHENKO, O. S., STARK, S. (2010). 75 Years After Likert: Thurstone Was Right! [online]. *Industrial and Organizational Psychology*, 3(4): 465–476. <https://doi.org/10.1111/J.1754-9434.2010.01273.X>.

EDWARDS, A. L. KENNEY, K. C. (1946). A comparison of the Thurstone and Likert techniques of attitude scale construction [online]. *Journal of Applied Psychology*, 30(1): 72–83. <https://doi.org/10.1037/H0062418>.

EDWARDS, A. L., KILPATRICK, F. P. (1948). Scale analysis and the measurement of social attitudes [online]. *Psychometrika*, 13(2): 99–114. <https://doi.org/10.1007/BF02289081>.

ESCHER, I. (2010). Pomiar kierunku i siły marketingowej postawy pracownika – kompromis pomiędzy teorią a praktyką marketingową [online]. *Acta Universitatis Nicolai Copernici Ekonomia*, 397(0): 159–174. <https://doi.org/10.12775/AUNC_ECON.2010.012>.

HOLBROOK, A., CHO, Y. I., JOHNSON, T. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties [online]. *Public Opinion Quarterly*, 70(4): 565–595. <https://doi.org/10.1093/poq/nfl027>.

Horizon Europe [online]. European Commission. [cit. 15.1.2022]. <https://ec.europa.eu/info/funding-tenders/find-funding/eu-funding-programmes/horizon-europe_en>.

LIPOVETSKY, S. (2007). Thurstone scaling in order statistics [online]. *Mathematical and Computer Modelling*, 45(7–8): 917–926. <https://doi.org/10.1016/J.MCM.2006.09.009>.

LIPOVETSKY, S., CONKLIN, W. M. (2004). Thurstone scaling via binary response regression [online]. *Statistical Methodology*, 1(1–2): 93–104. <https://doi.org/10.1016/J.STATMET.2004.04.001>.

MOSTELLER, F. (1951). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed [online]. *Psychometrika*, 16(2): 207–218. <https://doi.org/10.1007/BF02289116>.

SAFFIR, M. A. (1937). A comparative study of scales constructed by three psychophysical methods [online]. *Psychometrika*, 2(3): 179–198. <https://doi.org/10.1007/BF02288395>.

SINGLETON, R. A., STRAITS, B. C. (2018). *Approaches to Social Science Research.* Oxford UP, p. 635.

SOKOŁOWSKI, A. (1993). Propozycja testu podobieństwa struktur. *Przegląd Statystyczny*, 40(3–4): 295–301.

STADTHAGEN-GONZÁLEZ, H. et al. (2018). Using two-alternative forced choice tasks and Thurstone's law of comparative judgments for code-switching research [online]. *Linguistic Approaches to Bilingualism*, 8(1): 67–97. <https://doi.org/10.1075/LAB.16030.STA/CITE/REFWORKS>.

THURSTONE, L. L. (1927a). Psychological Analysis [online]. *The American Journal of Psychoanalysis*, 38(3): 368–389. <https://doi.org/10.3917/cohe.174.0156>.

THURSTONE, L. L. (1927b). The method of paired comparisons for social values [online]. *Journal of Abnormal and Social Psychology*, 21(4): 384–400. <https://doi.org/10.1037/H0065439>.

THURSTONE, L. L., CHAVE, E. J. (1929). *The measurement of attitude.* Chicago: The University of Chicago Press.

THURSTONE, L. L., JONES, L. V. (1957). The Rational Origin for Measuring Subjective Values [online]. *Journal of the American Statistical Association*, 52(280): 458–471. <https://doi.org/10.1080/01621459.1957.10501401>.

TSOGO, L., MASSON, M. H., BARDOT, A. (2000). Multidimensional Scaling Methods for Many-Object Sets: a Review [online]. *Multivariate Behavioral Research*, 35(3): 307–319. <https://doi.org/10.22004/ag.econ.135504>.

TSUKIDA, K., GUPTA, M. R. (2011). *How to Analyze Paired Comparison Data.* UWEE Technical Report, (206): 18.