Wiadomości Statystyczne. The Polish Statistician, 2024, *69*(11), 1–17 https://doi.org/10.59139/ws.2024.11.1
Studia metodologiczne / Methodological studies
Zgłoszony/submitted: 20.08.2023, poprawiony/revised: 21.07.2024, zaakceptowany/accepted: 2.09.2024

Proposed solutions to increase the reliability and goodness of fit of the Thurstone method

Joanna Dębicka,[a] Edyta Mazurek,[b] Katarzyna Ostasiewicz[c]

Abstract. The Thurstone method is a method of aggregation of preferences that leads to ordering objects at an interval scale, in contrast to other methods of aggregation, the application of which results in obtaining the ordinal scale only. However, we demonstrate that the interval scale obtained by means of the Thurstone method is not fully reliable, as it is strongly susceptible to the problem of irrelevant alternatives, i.e., the order of two objects might be dependent on whether there is or not a third, completely independent object in the studied collectivity.

The other issue examined in the paper is the formula that is to be minimised in the Thurstone method. Thurstone proposed a formula easy to be minimised with tabularised values of normal distribution. However, today's calculation powers of computers make it possible to minimise another formula, which allows a better fit of observed frequencies to the predicted ones, and makes it possible to avoid the problem with empirical frequencies close to 1.

The aim of the paper is to propose two improvements to the Thurstone method. The first one involves the application of the rule of thumb to the minimal spacing (in terms of standard deviations), below which the ordering of objects cannot be regarded as reliable, and the use of the method of empirical examining of the stability of order. The second of the proposed improvements consists in minimising a formula other than the original one, especially in the cases where empirical frequencies reach very high values.

Keywords: the Thurstone method, dependence on irrelevant alternatives, aggregation of preferences, ordering of objects

JEL: C18, C83, C46

Propozycja rozwiązań zwiększających wiarygodność i dobroć dopasowania w metodzie Thurstone'a

Streszczenie. Metoda Thurstone'a jest metodą agregacji preferencji, pozwalającą na porządkowanie obiektów na skali interwałowej, w odróżnieniu od innych metod agregacji, za pomocą których otrzymuje się jedynie skalę porządkową. Trudno jednak traktować skalę interwałową

[a] Uniwersytet Ekonomiczny we Wrocławiu, Wydział Ekonomii i Finansów, Polska / Wroclaw University of Economics and Business, Faculty of Economics and Finance, Poland.
ORCID: https://orcid.org/0000-0002-8905-2565. E-mail: joanna.debicka@ue.wroc.pl.

[b] Uniwersytet Ekonomiczny we Wrocławiu, Wydział Ekonomii i Finansów, Polska / Wroclaw University of Economics and Business, Faculty of Economics and Finance, Poland.
ORCID: https://orcid.org/0000-0001-7410-1638. E-mail: edyta.mazurek@ue.wroc.pl.

[c] Uniwersytet Ekonomiczny we Wrocławiu, Wydział Ekonomii i Finansów, Polska / Wroclaw University of Economics and Business, Faculty of Economics and Finance, Poland.
ORCID: https://orcid.org/0000-0002-0115-3696. Autor korespondencyjny / Corresponding author, e-mail: katarzyna.ostasiewicz@ue.wroc.pl.

uzyskaną przy użyciu metody Thurstone'a jako wiarygodną, ponieważ – jak pokazano w artykule – jest ona bardzo podatna na zależność od nieistotnych alternatyw: względna kolejność dwóch obiektów może zależeć od obecności innego, zupełnie niezależnego obiektu w badanym zbiorze.

W artykule poruszono również kwestię wyrażenia, które ma podlegać minimalizacji. Thurstone wskazał wyrażenie, które można łatwo zminimalizować za pomocą stablicowanych wartości rozkładu normalnego. Jednak wobec obecnych mocy obliczeniowych komputerów możliwe jest minimalizowanie innego wyrażenia, które daje lepszą dobroć dopasowania częstości obserwowanych do oczekiwanych oraz pozwala uniknąć problemu dotyczącego częstości empirycznych bliskich 1.

Artykuł ma na celu zaproponowanie dwóch ulepszeń metody Thurstone'a. Pierwszym jest stosowanie zgrubnej reguły dotyczącej progowej odległości między obiektami (w wielokrotnościach odchylenia standardowego), poniżej której uporządkowanie obiektów nie może być uważane za rzetelne, oraz metody empirycznego badania stabilności porządku obiektów. Drugie polega na minimalizowaniu wyrażenia innego niż oryginalne, w szczególności gdy wśród częstości empirycznych występują bardzo wysokie wartości.

Słowa kluczowe: metoda Thurstone'a, zależność od nieistotnych alternatyw, agregacja preferencji, porządkowanie obiektów

1. Introduction

The problem of ordering objects is an important issue in many fields of the public life; for example, public elections involve ordering 'objects', which in this case are the candidates. There are several other situations where ordering 'objects' is necessary, such as projects to be financed by the government, etc.

Many different methods of ordering objects have been investigated. One of them is the Thurstone method (Thurstone, 1927; Thurstone & Chave, 1929; Thurstone & Jones, 1957), which is regarded useful not only for ordering objects, but also for finding relative spaces between them. While in e.g. a presidential election it is only important who wins (i.e. who obtains the 'first rank'), in other situations – for instance, distributing funds – it might be necessary to know which subsequent results are very close to one another and which are largely distanced.

The Thurstone method has been widely used in various contexts (e.g. Dragonetti et al., 2017; Heldsinger & Humphry, 2010; Kanda, 2002; Krabbe, 2008; Orbán-Mihálykó et al., 2022; Temesi et al., 2023; van Wijk & Hoogstraten, 2004; Woods et al., 2010), and its properties have been investigated intensively (e.g. Handley, 2001; Lipovetsky, 2007; Lipovetsky & Conklin, 2004; Vojnovic & Yun, 2016; Yellott, Jr., 1980). However, there seems to be few studies with respect to empirical results of applying the Thurstone method to real data.

We already discussed the subject of testing the Thurstone method (Dębicka et al., 2023). We showed that although a very smart test was developed for the Thurstone model (Mosteller, 1951), if it is applied to a sample consisting of 100 or more items, almost always the results will turn out statistically insignificant. In this paper, we show that although the Thurstone method is touted as finding the interval scale, this scale is prone to change when one or more objects are removed from the set. This is called 'dependence on irrelevant alternatives', and obviously it is not a desirable property of a method.

We begin by having a closer look at the historical conditioning of the Thurstone method, i.e. the formulae to be minimised in order to obtain the estimates of modes of distributions. These formulae are still used nowadays, even though these days, aided by advanced computational powers of contemporary computers, scientists could come up with far more effective solutions.

We continue with the investigation of the stability of the scale obtained with the Thurstone method on the basis of two sets of data, i.e. the original Thurstone data (1927) and the data from our own survey (Dębicka et al., 2022). We are empirically checking the stability of the ordering of two items when changes in the set of objects and in the set of respondents occur.

The aim of our paper is twofold. Firstly, we propose the rule of thumb relating to the minimal spacing (in terms of standard deviations), below which the ordering of objects cannot be regarded as reliable, and the methods of empirical examining stability of order. Secondly, we indicate the possibility of minimising a formula other than the original one, especially in the cases where very high empirical frequencies occur.

2. The Thurstone method

According to the most common and simplest presentation of the Thurstone method, we have *n* objects, and each of them has some objective position. However, individuals can make errors comparing any pair of these objects and those errors are distributed normally (in the simplest case, with the same standard deviation σ for each pair). The greater the difference between the objects (given in units of σ), the smaller the probability that the respondent will misidentify a smaller object for a greater one. Let us express position (modes), l_k, in terms of multiplicity, x_k, of a common standard deviation σ: $l_k = x_k \cdot \sigma$. For objects i and j with positions equal $l_i = x_i \cdot \sigma$ and $l_j = x_j \cdot \sigma$, $x_j > x_i$, this probability of misidentification is equal to:

$$1 - p_{ji} = P(l_i > l_j) = F(x_i - x_j) = 1 - F(x_j - x_i),$$

where p_{ji} denotes the probability of proper identification of the order of objects *i* and *j*, and *F* is the cumulative distribution of the standard normal random variable.

If we had an infinite number of observations, the observed frequency of respondents who correctly identified the order of objects would be equal exactly to the theoretical probability:

$$\omega_{ji} = p_{ji},$$

where ω_{ji} denotes the observed frequency of respondents stating that object j is greater than i.

However, due to the inevitable limitations of the empirical study, in general, the observed frequency will deviate from real probabilities. Thus, our task is to estimate the real probabilities and real positions of the objects on the basis of the observations. To this end, the expression:

$$\mathcal{F}_1(\{x_i\}) = \sum_{j>i}\left[F^{-1}(\omega_{ji}) - (x_j - x_i)\right]^2, \tag{1}$$

is minimised with respect to x_i.

As zero on the obtained scale is arbitrary, one may conveniently rescale $\{x_i\}$ into a $[0, 1]$ interval:

$$x_i' = \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}.$$

In this way, we obtain not only the order of objects, but also their relative spacing, that is, not an ordinal scale, but an interval one. This is the advantage of the Thurstone method over simpler ones, e.g. average ranks method.

3. Problem with infinite inverse distribution

In the original Thurstone method, the object of minimisation is formula (1), and all the existing programs that have implemented the Thurstone method apply this approach. However, some problems with infinities may appear, as $F^{-1}(1) = \infty$ and $F^{-1}(0) = -\infty$. It is not highly improbable to obtain such values, if one of the objects in the set is clearly greater than all the others. In such a case, one may remove such an object from the whole set. Another way is to treat the problem purely numerically, e.g. to convert the empirical frequency 1 into 0.9999 or some other quantity close to 1, but not generating infinities. For example, the Statistica program replaces frequency 1 with 0.9999, but this value is completely arbitrary – why 0.9999 and not 0.999 or 0.99999? In other programs, e.g. SPSS or Stata, the Thurstone method is not implemented – the user does the analysis employing the general tools available in these programs, and has to manage him- or herself the problem with frequencies 0% or 100%. However, such a technical 'trick' seems to be artificial and arbitrary, and the final results for the whole scale may depend on the actual value adopted, as it is the very essence of the Thurstone method (i.e. that the position of object *i* may depend on frequency $\omega_{jk}, j, k \neq i$). Let us say, we have a matrix of empirical frequencies as given in Table 1.

Table 1. Sample values illustrating the problem of approximating the 100% frequency

j \ i	O1	O2	O3
O1	.	1	0.35
O2	0	.	0.99
O3	0.65	0.01	.

Note. The quantities in table (ω_{ij}) denote empirical frequencies of respondents preferring object i over object j.
Source: authors' calculations based on fictious data.

If one converted 1 into e.g. 0.999999, the scaled result would be 0, 1, 0.93 (for O1, O2 and O3, respectively). If one decided for two decimal parts less, i.e. 0.9999, it would be 0, 0.90, 1, i.e. O2 and O3 would change their relative positions. If one used inverse PDF equal to 3.05 (as Krabbe, 2008, did), it would be 0, 0.75, 1. This obviously is an artificial example, but it shows the problem with empirical frequencies equal to 1 or 0.

However, the formula that Thurstone decided to minimise in his method might have a historical background. An alternative formula to be minimised, which still complies with the general idea of minimising the deviations of the empirical data from the predicted values is:

$$\mathcal{F}_2 = \sum_{j>i}\left[\omega_{ji} - F\left(x_j - x_i\right)\right]^2. \tag{2}$$

Please note that the same authors who minimise the original Thurstone expression use the following quantity as a criterion of goodness of fit (cf. e.g. Krabbe, 2008; Lipovetsky, 2007; Mosteller, 1951) and the same expressions underly the test (Mosteller, 1951):

$$\chi^2 = \frac{\sum_{ij}\left(p_{ij} - \omega_{ij}\right)^2}{\sigma_{ij}^2}. \tag{3}$$

It is obvious that by minimising (2), one would obtain a better fit. So, why use (1)? The historical reason is not likely that in the 1920s, minimising the latter was completely impossible, as computers were not available. Using (1), one might apply tables (as Thurstone did). Please note that by today's standards of accuracy, there are some errors in Thurstone's results (1927) due to the erroneous rounding. However, it is still tempting to use (1), as one obtains analytical formulae this way, but the problem with infinities and worse goodness of fit remains.

If the data perfectly fitted the predictions, both approaches – minimising (1) and (2) – would yield the same results, which is obvious, as there would exist such quantities as to make both (1) and (2) equal exactly 0. However, if there were some deviations, the results yielded by both methods would slightly differ.

In his seminal paper, Thurstone (1927) presented results of the analysis of 19 crimes using his own method to find ordering of these crimes. We repeated the analysis of the same data by means of formula (2), instead of the original one, to compare the results. They are presented in Table 2.

Table 2. Comparison of Thurstone's original results with the scale obtained by minimising (2)

Type of crime	Original Thurstone (1)	Minimising (2)
Abortion	0.69326	0.74232
Adultery	0.64146	0.68019
Arson	0.61714	0.64457
Assault	0.45010	0.49320
Bootlegging	0.31527	0.35629
Burglary	0.46059	0.50698
Counterfeiting	0.49876	0.53702
Embezzlement	0.50608	0.55089
Forgery	0.47691	0.51623
Homicide	0.96322	0.99414
Kidnapping	0.67115	0.68534
Larceny	0.40448	0.44478
Libel	0.34304	0.38178
Perjury	0.51172	0.53799
Rape	1	1
Receiving stolen goods	0.30464	0.34566
Seduction	0.69368	0.72590
Smuggling	0.33652	0.37149
Vagrancy	0	0

Source: authors' calculations based on data from Thurstone (1927).

As one can see, no object changed its relative position with respect to the others. Still, spacing between the objects varied in some cases even by as much as over 10%. If we are interested in proper spacing – which is usually presented as the merit of the Thurstone method – should not it rather be these distances that fit the data better?

4. Dependence on irrelevant alternatives

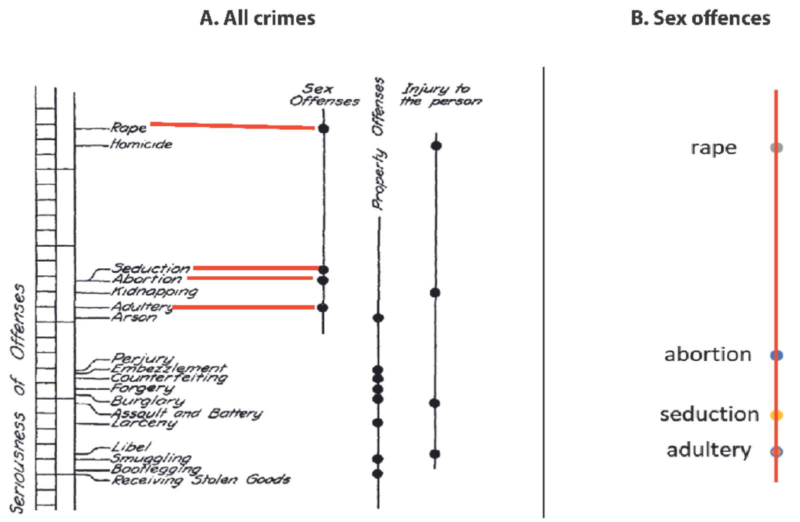
The Thurstone method is an aggregation procedure. It is rather impossible that there is some objective, set scale of crimes, and the fact there are differences between particular people's ordering of crimes results from random errors. More likely, each person has his or her own, subjective scale. Therefore, the task of researchers is to construct an aggregated scale for the whole population.

On the other hand, according to Arrow's theorem, there is no method of collective choice that would meet some reasonable assumptions (e.g. dictatorship is bad) and would be independent from irrelevant alternatives (Arrow, 1951). This applies to the Thurstone method (it being the method of preference aggregation) as well – the ordering of preferences done by means of this method would be, at least theoretically, dependent on irrelevant alternatives. However, one might ask here whether this problem is frequently encountered in empirical data. If it is, another question would have to be posed, namely how to interpret the ordering of preferences with such a shortcoming.

Thurstone (1927) investigated 19 crimes – the respondents had to compare each pair of them. Thus, by applying his method, he obtained a linear ordering of those crimes. Subsequently, he was able to present an order within the subgroups of crimes, dividing them into 'sex offences', 'property offences' and 'injury to the person'.

The disturbing fact is that if sexual crimes were just taken off the general ordering (see Figure 1a), their relative order would be quite different from the order obtained with data limited only to these pairs of crimes where both crimes belonged to the 'sexual offences' subgroup (cf. Figure 1b). Therefore, to construct 1b, we dealt with a restricted matrix of frequencies, i.e. only 4 x 4 (for those four crimes of our special interest). But we can see that in this approach, 'seduction' and 'abortion' changed their relative positions and the spacing between them changed significantly. As the respondents were supposed to compare each pair independently, this result is troublesome to say the least. If we were to decide on the basis of this result which of the sexual crimes to punish most harshly or on preventing which of them to spend most money or effort, we would have trouble deciding which result should be treated as the reliable one. The fact that it is possible the result of aggregating preferences depends on irrelevant objects which might (but do not have to) be taken into consideration in the whole procedure is indeed troublesome from the scientific point of view.

Figure 1. Ordering all crimes and sex offences



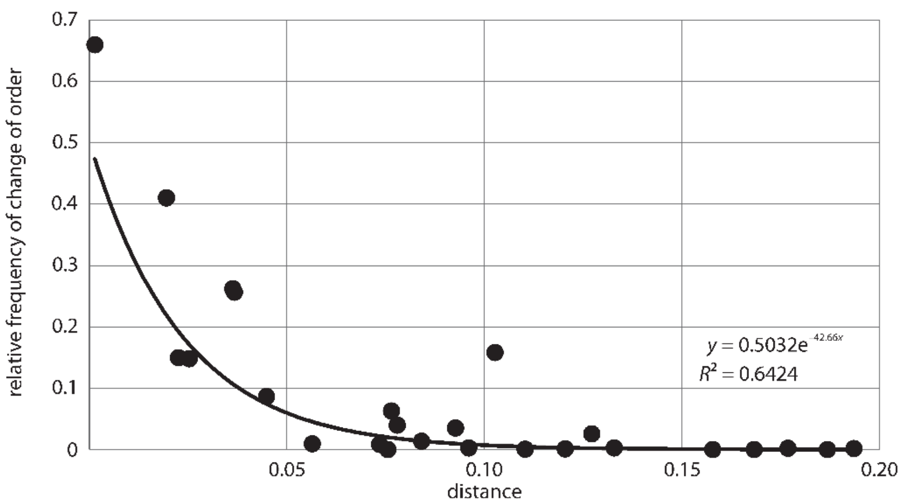
Source: A – Thurstone (1927), B – authors' work.

To investigate the deepness of the problem, we performed a detailed and complete analysis of the dataset presented by Thurstone. We examined every possible combination of crimes, i.e. starting from the set of two crimes, we took into account all the possible pairs, finally reaching all combinations for the 18 out of all 19 crimes. The research question was: are there any further changes of order among the crimes, like this change which we observed within the set of sexual crimes? Which crimes are most prone to this 'instability' of position?

Table 3 (p. 10) shows all the obtained inconsistencies. We treated the ordering of all the 19 crimes as a reference point, and listed all the deviations from the relative order in all possible subsets (combinations of crimes). Table 3 presents the relative frequencies of these changes (each pair of objects appears together in the total number of 131,070 subsets). The objects are ordered according to the overall Thurstone scale, which means that the closer to the diagonal a given cell is, the closer the crimes from a given pair are to each other on the Thurstone scale (constructed for the whole set of crimes). The most frequently observed alteration was the relative change of positions of the crimes neighbouring on the overall scale – nearly all these pairs changed at least in some subsets – but also in relation to the second nearest neighbours and even third nearest neighbours. In fact, the propensity to changing the position depends not on the relative ranks of objects, but rather on the distance between them.

Table 4 (p. 11) presents the distances between all the pairs of objects (measured in units of standard deviation), with those cells which show distances lower than 0.2 highlighted. It is striking how these highlighted cells coincide with non-zero values in Table 3. Moreover, if we gather all non-zero change-of-order relative frequencies and corresponding distances, the relationship between those quantities is strong, as pictured in Figure 2.

Figure 2. Relative frequencies of change of order of pairs of objects vs the distances between them as obtained by the Thurstone procedure



Source: authors' work based on data from Thurstone (1927) and authors' own calculations.

We repeated the same procedure for data obtained in our own research (Dębicka et al., 2022), with the set of crimes different in some parts from Thurstone's original set, in order to fit our reality better. The crimes were arranged in the following order: violent rape, assault with a severe body injury, pedophilic acts, domestic violence, threats, murder, defamation (slander), harassment, kidnapping for ransom and identity theft. Our results are presented in Tables 5 and 6 (for frequencies of changes and distances, respectively, p. 12). One can see that these results were much more stable than Thurstone's original results when we analysed their dependence on irrelevant alternatives. There are only two pairs whose relative order might be questioned – and these are the same two pairs which are separated by the smallest distances on the overall Thurstone scale.

Table 3. Relative frequencies of exchanging order based on Thurstone's data

Type of crime	O19	O16	O5	O18	O13	O12	O4	O6	O9	O7	O8	O14	O3	O2	O11	O1	O17	O10	O15
O19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O16	0	0	0.256	0.001	0.003	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O5	0	0.256	0	0.009	0.003	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O18	0	0.001	0.009	0	0.150	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O13	0	0.003	0.003	0.150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O12	0	0	0	0	0	0	4E–5	0	0	0	0	0	0	0	0	0	0	0	0
O4	0	0	0	0	0	4E–5	0	0.262	0.035	E–04	0.002	0	0	0	0	0	0	0	0
O6	0	0	0	0	0	0	0.262	0	0.010	0	0	0.002	0	0	0	0	0	0	0
O9	0	0	0	0	0	0	0.035	0.009	0	2E–4	0	0.001	0	0	0	0	0	0	0
O7	0	0	0	0	0	0	E–4	0	2E–4	0	0.148	0.086	0	0	0	0	0	0	0
O8	0	0	0	0	0	0	0.002	0	0	0.148	0	0.410	0	0	0	0	0	0	0
O14	0	0	0	0	0	0	0	0.002	0.001	0.086	0.410	0	0	0	0	0	0	0	0
O3	0	0	0	0	0	0	0	0	0	0	0	0	0	0.014	2E–5	0	0	0	0
O2	0	0	0	0	0	0	0	0	0	0	0	0	0.014	0	0.158	0	0	0	0
O11	0	0	0	0	0	0	0	0	0	0	0	0	2E–5	0.158183	0	0.063	0.040	0	0
O1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.063	0	0.660	0	0
O17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.040	0.660	0	0	0
O10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.026
O15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.026	0

Note. O1–O19 stand for the 19 crimes. Values greater than 0 are highlighted.

Source: authors' calculations based on data from Thurstone (1927).

Table 4. Distances between crimes based on Thurstone's data

Type of crime	O19	O16	O5	O18	O13	O12	O4	O6	O9	O7	O8	O14	O3	O2	O11	O1	O17	O10	O15
O19	0.00	1.05	1.09	1.16	1.19	1.40	1.56	1.59	1.65	1.72	1.75	1.77	2.13	2.22	2.32	2.40	2.40	3.33	3.46
O16	1.05	0.00	0.04	0.11	0.13	0.35	0.50	0.54	0.60	0.67	0.70	0.72	1.08	1.16	1.27	1.34	1.35	2.28	2.40
O5	1.09	0.04	0.00	0.07	0.10	0.31	0.47	0.50	0.56	0.63	0.66	0.68	1.04	1.13	1.23	1.31	1.31	2.24	2.37
O18	1.16	0.11	0.07	0.00	0.02	0.23	0.39	0.43	0.49	0.56	0.59	0.61	0.97	1.05	1.16	1.23	1.23	2.17	2.29
O13	1.19	0.13	0.10	0.02	0.00	0.21	0.37	0.41	0.46	0.54	0.56	0.58	0.95	1.03	1.13	1.21	1.21	2.14	2.27
O12	1.40	0.35	0.31	0.23	0.21	0.00	0.16	0.19	0.25	0.33	0.35	0.37	0.74	0.82	0.92	1.00	1.00	1.93	2.06
O4	1.56	0.50	0.47	0.39	0.37	0.16	0.00	0.04	0.09	0.17	0.19	0.21	0.58	0.66	0.76	0.84	0.84	1.77	1.90
O6	1.59	0.54	0.50	0.43	0.41	0.19	0.04	0.00	0.06	0.13	0.16	0.18	0.54	0.63	0.73	0.80	0.81	1.74	1.87
O9	1.65	0.60	0.56	0.49	0.46	0.25	0.09	0.06	0.00	0.08	0.10	0.12	0.48	0.57	0.67	0.75	0.75	1.68	1.81
O7	1.72	0.67	0.63	0.56	0.54	0.33	0.17	0.13	0.08	0.00	0.03	0.04	0.41	0.49	0.60	0.67	0.67	1.61	1.73
O8	1.75	0.70	0.66	0.59	0.56	0.35	0.19	0.16	0.10	0.03	0.00	0.02	0.38	0.47	0.57	0.65	0.65	1.58	1.71
O14	1.77	0.72	0.68	0.61	0.58	0.37	0.21	0.18	0.12	0.04	0.02	0.00	0.36	0.45	0.55	0.63	0.63	1.56	1.69
O3	2.13	1.08	1.04	0.97	0.95	0.74	0.58	0.54	0.48	0.41	0.38	0.36	0.00	0.08	0.19	0.26	0.26	1.20	1.32
O2	2.22	1.16	1.13	1.05	1.03	0.82	0.66	0.63	0.57	0.49	0.47	0.45	0.08	0.00	0.10	0.18	0.18	1.11	1.24
O11	2.32	1.27	1.23	1.16	1.13	0.92	0.76	0.73	0.67	0.60	0.57	0.55	0.19	0.10	0.00	0.08	0.08	1.01	1.14
O1	2.40	1.34	1.31	1.23	1.21	1.00	0.84	0.80	0.75	0.67	0.65	0.63	0.26	0.18	0.08	0.00	0.00	0.93	1.06
O17	2.40	1.35	1.31	1.23	1.21	1.00	0.84	0.81	0.75	0.67	0.65	0.63	0.26	0.18	0.08	0.00	0.00	0.93	1.06
O10	3.33	2.28	2.24	2.17	2.14	1.93	1.77	1.74	1.68	1.61	1.58	1.56	1.20	1.11	1.01	0.93	0.93	0.00	0.13
O15	3.46	2.40	2.37	2.29	2.27	2.06	1.90	1.87	1.81	1.73	1.71	1.69	1.32	1.24	1.14	1.06	1.06	0.13	0.00

Note. O1–O19 stand for the 19 crimes. Values smaller than 0.2 are highlighted.

Source: authors' calculations based on data from Thurstone (1927).

Table 5. Relative frequencies of exchanging order for the Thurstone method applied to 10 crimes from the authors' survey

Type of crime	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0.000	0.000	0.753	0.000	0.000	0.000	0.000	0.000	0.000	0.000
O2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
O3	0.753	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
O4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000
O5	0.000	0.000	0.000	0.000	0.000	0.000	0.110	0.000	0.000	0.000
O6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
O7	0.000	0.000	0.000	0.000	0.110	0.000	0.000	0.000	0.000	0.000
O8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
O9	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000
O10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note. O1–O10 stand for the 10 crimes. Values on the diagonal and greater than 0.1 are highlighted.
Source: authors' calculations based on the data from their own survey.

Table 6. Distances between the 10 crimes, obtained by the Thurstone method applied to data from authors' survey

Type of crime	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0.00	0.37	0.02	0.82	2.41	0.25	2.47	1.95	1.05	1.59
O2	0.37	0.00	0.35	0.45	2.04	0.63	2.09	1.57	0.68	1.22
O3	0.02	0.35	0.00	0.80	2.39	0.28	2.44	1.92	1.03	1.57
O4	0.82	0.45	0.80	0.00	1.59	1.08	1.64	1.12	0.23	0.77
O5	2.41	2.04	2.39	1.59	0.00	2.67	0.05	0.47	1.36	0.82
O6	0.25	0.63	0.28	1.08	2.67	0.00	2.72	2.20	1.30	1.85
O7	2.47	2.09	2.44	1.64	0.05	2.72	0.00	0.52	1.42	0.87
O8	1.95	1.57	1.92	1.12	0.47	2.20	0.52	0.00	0.90	0.35
O9	1.05	0.68	1.03	0.23	1.36	1.30	1.42	0.90	0.00	0.54
O10	1.59	1.22	1.57	0.77	0.82	1.85	0.87	0.35	0.54	0.00

Note. O1–O10 stand for the 10 crimes. Values that are equal or smaller than 0.05 are highlighted.
Source: authors' calculations based on data from their own survey.

It is obvious that the smaller the distance between the objects, the higher the probability of misidentifying them (i.e. the probability that the order of two objects will be reversed). If the distance between the expected values is x (in terms of standard deviation), the probability of a single person misidentifying them amounts to $p(x) = 1 - F(x)$. For a large number of respondents, n, the probability that the whole group would misidentify the objects can be approximated by

$$1 - F\left(\frac{0.5 - p(x)}{\sqrt{\frac{p(x)\big(1 - p(x)\big)}{n}}}\right).$$

To continue the example from Tables 5 and 6, one may calculate those probabilities for the most problematic pairs of objects, as presented in Table 7.

Table 7. Probabilities of the misidentifications of objects for: a single person, 220 respondents and 1,000 respondents

Specification	O1, O6	O3, O1	O2, O3	O4, O2	O9, O4	O10, O9	O8, O10	O5, O8	O7, O5
Distance	0.253	0.025	0.350	0.449	0.227	0.544	0.353	0.467	0.054
$N = 1$	0.400	0.490	0.363	0.327	0.410	0.293	0.362	0.320	0.479
$N = 220$	0.001	0.385	10^{-5}	10^{-8}	0.003	10^{-12}	10^{-5}	10^{-9}	0.262
$N = 1,000$	10^{-11}	0.267	0	0	10^{-9}	0	0	0	0.087

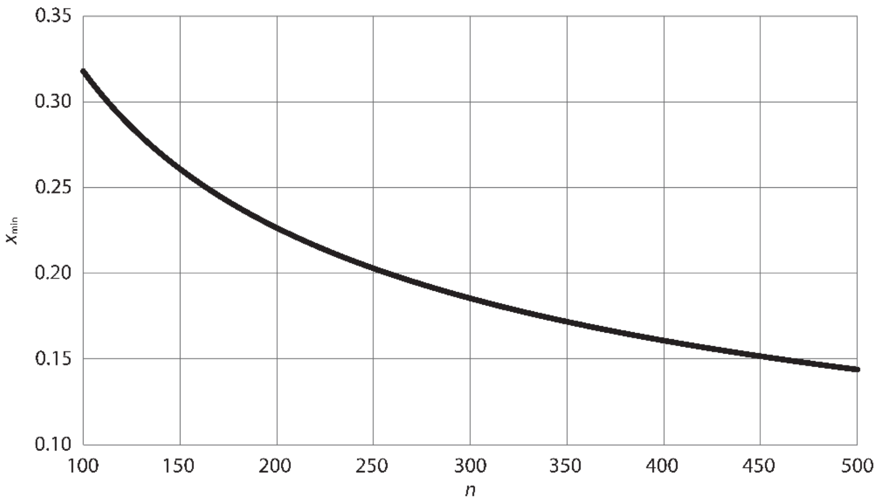
Note. The numbers of objects in the caption of the table correspond to O1–O10 in Tables 5–6.
Source: authors' calculations based on data from their own survey.

In the '$N = 1$' row of Table 7, the probabilities of an object misidentification for a single person are presented, in the '$N = 220$' row, the probabilities that the number of misidentifications will be greater than half of the trials for $N = 220$ (the actual number of respondents in our survey), and in the '$N = 1,000$' row, the probabilities of misidentification for a much greater number of respondents, i.e. $N = 1,000$. One might observe that for $N = 220$, the probability of a group misidentification of such close objects is very high, namely 39% for the first pair and 26% for the second one. And even if we increase the number of respondents to 1,000, this probability will still be high: 27% and 9%, respectively. For comparison, in the case of $N = 220$, the remaining probabilities were all less than 0.35% (the greatest one), and for $N = 1,000$, they were of order 10^{-9} at most.

Thus, if we require relatively small probability of misidentification, e.g. 0.005, we can calculate the limiting distance between the objects below which we could not distinguish between them on the basis of an n-element sample (see Figure 3).

Figure 3 presents the limiting value of a distance between the objects for a given n (between 100 and 500), below which the probability of misidentification will be higher than 0.005. This limiting value coincides to a large degree with the values for which we can observe changes of relative positions depending on the subset. Thus, the probability of misidentification below 0.005 may need the rule of thumb to decide which objects cannot be regard as properly ordered within the scope of the Thurstone method.

Figure 3. The limiting value of the distance between the objects for a given *n*, below which the probability of misidentification will be higher than 0.005



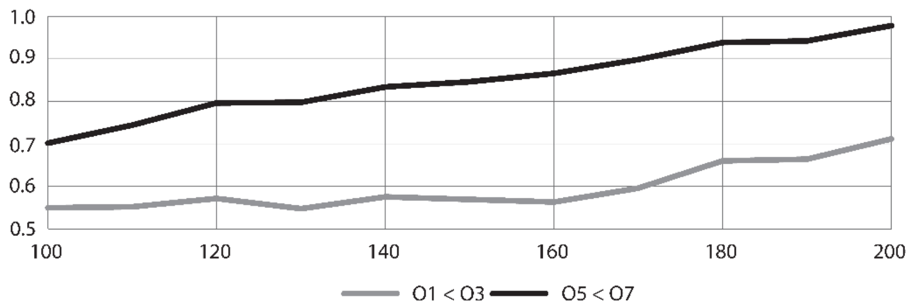
Source: authors' work based on their calculations.

5. Respondents-subsets analysis

In this part of the sensitivity analysis, we checked what the results would be if not all respondents were surveyed. To this end, we chose at random 1,000 different 100-element sets of respondents (each set chosen at random) and found the resulting order of objects. As for some pairs of objects the answers were almost uniform (the relative frequencies approaching 1 or 0), in order to avoid the problem with infinite inverse cumulative distribution function we introduced one 'artificial person', whose answers were always against the general tendency, and who was always included in the set of other respondents, chosen at random. It turned out that the pairs whose order could be changed (as compared to the set of all respondents) were O1, O3 and O5, O7 ones. In 54% of the rankings, O1 was lower than O3, and in 74% pairs, O5 was lower than O7. Thus, it appears that a 100-element sample is not sufficient to fix the order of O1, O3 and O5, O7. However, the remaining objects in all the 1,000 cases had exactly the same order.

One may suggest that a 100-element sample is not large enough. So the question is whether by increasing the sample size (but keeping it below 219, to be able to choose many different samples of a given size), we would obtain more stable results. Figure 4 presents the fractions of orders O1 < O3 and O5 < O7 for sample sizes between 100 to 200.

Figure 4. The fractions of orders O1 < O3 and O5 < O7 for sample sizes from 100 to 200 based on data from authors' survey



Source: authors' work.

What can be seen is that although the pair O5, O7 is more quickly dominated by the order O5 < O7 the order still happens to be reversed for the 200-element sample. As regards the O1, O3 pair, this phenomenon is even more striking.

The conclusion of these two kinds of a 'sensitivity analysis' might be the following. Although we do not necessarily assume that the 'objective' position of objects form a linear order, we should expect that the result of aggregating preferences would give some stable results, not depending on slight changes in the procedure, e.g. adding one or two other objects, or removing a few from the sample. Thus, the relative order of O1, O3 and O5, O7 cannot be regarded as reliable within this research. It might be the case that the expansion of the study to larger groups of respondents would yield more stable results. On the other hand, we cannot be sure that adding some other objects to the analysed set would not destabilise the order of remaining objects (which turned out to be stable in this study). Also, we cannot be sure that expanding the sample by more respondents would cause the results to remain unchanged. However, the detailed sensitivity analysis at least corroborated the presumption that the order of O2, O4, O6, O8, O9 and O10 would remain stable (as was the case in our study).

6. Conclusions

The Thurstone method is often believed to be capable of establishing an interval scale of a set of objects, in contrast to e.g. the average rank method.

However, it is hard to consider the distances between the objects that are smaller than approximately 0.2 of the standard deviation of normal distributions (underlying the Thurstone method) as valid. Most of such pairs of objects are unstable when the set of other objects they are examined with changes. This is called 'dependence on irrelevant alternatives', and obviously it is not a desirable property for any method.

Although we know that in general, none of the methods of aggregating preferences is free of this flaw, in practice some of them are more prone to be dependent on irrelevant alternatives than others. Our empirical study which used Thurstone's original data and the data we obtained by repeating Thurstone's procedure a century later demonstrates that the Thurstone method is prone to the dependency on irrelevant alternatives to a relatively large degree. On the other hand, no empirical study can prove or disprove a general property. Therefore what we suggest is not to relay too much on a method and to check its resilience every time. Pairs of objects which frequently exchange their relative ranks should be thus regarded as undistinguishable for all practical purposes. Treating one of the objects from such a pair as larger solely on the basis of it being larger in a given set of objects depends too much on the arbitrary decision which objects were and which were not included in the study.

Another observation related to the Thurstone method regards the choice of the formula to be minimised. The original formula was chosen by Thurstone in the 1920s most probably due its relative calculational simplicity. However, this historical formula is still used nowadays, which causes problems with high empirical frequencies close to 1. We suggest that researchers nowadays might apply formula (2) to the minimising procedure, which would automatically ensure a better fit of the predicted frequencies to the observed ones.

References

Arrow, K. J. (1951). *Social Choice and Individual Values.* Wiley.

Dębicka, J., Mazurek, E., & Ostasiewicz, K. (2022). Methodological Aspects of Measuring Preferences Using the Rank and Thurstone Scale. *Statistika*, *102*(3), 236–248. https://doi.org/10.54694/stat.2022.5.

Dębicka, J., Mazurek, E., & Ostasiewicz, K. (2023). *Thurstone model – a model or a procedure?* [manuscript submitted for publication].

Dragonetti, R., Ponticorvo, M., Dolce, P., Di Filippo, S., & Mercogliano, F. (2017). Pairwise comparison psychoacoustic test on the noise emitted by DC electrical motors. *Applied Acoustics*, *119*, 108–118. https://doi.org/10.1016/j.apacoust.2016.12.016.

Handley, J. C. (2001). *Comparative Analysis of Bradley-Terry and Thurstone-Mosteller Paired Comparison Models for Image Quality Assessment*. PICS: Image Processing, Image Quality, Image Capture, Systems Conference, Montreal. https://www.imaging.org/common/uploaded%20files/pdfs/Papers/2001/PICS-0-251/4604.pdf.

Heldsinger, S., & Humphry, S. (2010). Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments. *The Australian Educational Researcher*, *37*(2), 1–19. https://doi.org/10.1007/BF03216919.

Kanda, T. (2002). Classification of menus on home dining tables based upon human meal Kansei. *Kansei Engineering International*, *3*(4), 9–14. https://doi.org/10.5057/kei.3.4_9.

Krabbe, P. F. (2008). Thurstone Scaling as a Measurement Method to Quantify Subjective Health Outcomes. *Medical Care*, *46*(4), 357–365. https://doi.org/10.1097/MLR.0b013e31815ceca9.

Lipovetsky, S. (2007). Thurstone scaling in order statistics. *Mathematical and Computer Modelling*, *45*(7–8), 917–926. https://doi.org/10.1016/j.m cm.2006.09.009.

Lipovetsky, S., & Conklin, W. M. (2004). Thurstone scaling via binary response regression. *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. *Psychometrika*, *16*(2), 207–218. https://doi.org/10.1007/BF02289116.

Orbán-Mihálykó, É., Mihálykó, C., & Gyarmati, L. (2022). Application of the Generalized Thurstone Method for Evaluations of Sports Tournaments' Results. *Knowledge*, *2*(1), 157–166. https://doi.org/10.3390/knowledge2010009.

Temesi, J., Szádoczki, Z., & Bozóki, S. (2023). Incomplete pairwise comparison matrices: Ranking top women tennis players. *Journal of the Operational Research Society*, *75*(1), 145–157. https://doi.org/10.1080/01605682.2023.2180447.

Thurstone, L. L. (1927). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, *21*(4), 384–400. https://psycnet.apa.org/doi/10.1037/h0065439.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. The University of Chicago Press.

Thurstone, L. L., & Jones, L. V. (1957). The Rational Origin for Measuring Subjective Values. *Journal of the American Statistical Association*, *52*(280), 458–471. https://psycnet.apa.org/doi /10.2307/2281694.

Vojnovic, M., & Yun, S.-Y. (2016). Parameter Estimation for Generalized Thurstone Choice Models. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (vol. 48, pp. 498–506). JMLR.org.

van Wijk, A., & Hoogstraten, J. (2004). Paired comparisons of sensory pain adjectives. *European Journal of Pain*, *8*(4), 293–297. https://doi.org/10.1016/j.ejpain.2003.10.002.

Woods, R. L., Satgunam, P., Bronstad, M., & Peli, E. (2010). Statistical analysis of subjective preferences for video enhancement. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human Vision and Electronic Imaging XV* (vol. 7527, pp. 98–107). SPIE.

Yellott, Jr., J. I. (1980). Generalized Thurstone models for ranking: equivalence and reversibility. *Journal of Mathematical Psychology*, *22*(1), 48–69. https://doi.org/10.1016/0022-2496(80)90046-2.