ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Silesian Statistical Review

2024, nr 22(28) ISSN 2449-9765

DOI: 10.15611/sps.2024.22.06

Thurstone Model: Analysing Survey Data and Ranking Stimuli – Model or Procedure?

Katarzyna Ostasiewicz

Wroclaw University of Economics and Business

e-mail: katarzyna.ostasiewicz@ue.wroc.pl

ORCID: 0000-0002-0115-3696

Joanna Dębicka

Wroclaw University of Economics and Business

e-mail: joanna.debicka@ue.wroc.pl

ORCID: 0000-0002-8905-2565

Edyta Mazurek

Wroclaw University of Economics and Business

e-mail: edyta.mazurek@ue.wroc.pl

ORCID: 0000-0001-7410-1638

Quote as: Ostasiewicz, K., Dębicka, J., & Mazurek, E. (2024). Thurstone Model: Analysing Survey Data and Ranking Stimuli – Model or Procedure? *Silesian Statistical Review*, *22*(28), 50-61.

JEL: C10, C44, D01

Abstract: This study explores the suitability of the Thurstone model in preference studies through empirical validation. While the Thurstone model has been widely utilised for establishing linear orders among stimuli, its appropriateness has come under scrutiny. The research proposes a novel empirical validation test to assess the model's applicability, particularly concerning large sample sizes. Findings suggest limitations in using the Thurstone model with extensive samples. Consequently, the paper offers insights for researchers relying on this model for stimulus ranking and provides guidelines for verifying its suitability. Moreover, it discusses alternative pathways for addressing poor model fit, including incorporating different standard deviations and treating the Thurstone method not as a definitive model but as a procedure for ordering objects.

Additionally, the study compares the Thurstone method with the average rank method, highlighting the advantages and challenges of each approach. Overall, this research contributes to understanding the complexities of preference studies and provides valuable insights into refining methodological approaches in this domain.

Keywords: Thurstone model, preference studies, empirical validation, stimulus ranking

1. Introduction

The Thurstone model was first proposed in the early 20th century (Thurstone, 1927a, 1927b; Thurstone & Chave, 1929; Thurstone & Jones, 1957) and has since been frequently utilised to determine a linear order within a set of stimuli under investigation. In this approach, respondents are asked to judge which pair of stimuli is more excellent, and the results are presented as a matrix of frequencies. Achieving a linear order has found wide application in classical social choice (De Borda, 1781), information retrieval (Dwork et al., 2001), and recommendation systems (Baltrunas et al., 2010). Ranking methods and paired comparisons play a crucial role in measuring preferences, attitudes, and values. Rank aggregation methods have also been extensively employed in marketing and advertising research, as well as in applied psychology. In recent years, rank aggregation methods have emerged as a vital tool for combining information from various Internet search engines (Lin, 2010). Determining the order of a set of objects based on pairwise comparisons is much simpler than the traditional ranking of the entire set of objects.

Moreover, the comparative format used in ranking and paired comparison tasks can significantly mitigate the influence of uniform response biases typically associated with rating scales (Maydeu-Olivares & Brown, 2010). Maydeu-Olivares and Böckenholt (2008) present a list of the top 10 reasons for using Thurstonian models. Notably, it is easy for respondents, inconsistencies can be modelled, the validity of inferences can be tested and nested, and crossed-sampling structures can be incorporated. For these reasons, the Thurstone model is frequently employed in practice. The simplest form of the model assumes that each stimulus follows a normal distribution with the same standard deviation and that all correlations among stimuli are equal. Then, the method of least squares is applied to minimise the deviation of the inverse CDFs of observed frequencies, and the estimated central positions of linearly ordered normal distributions:

$$\sum_{ij}\left(F^{-1}(\omega_{ij})-(d_j-d_i)\right)^2\to\min, \tag{1}$$

where $F^{-1}(\omega_{ij})$ denotes the inverse cumulative distribution of the standard normal random variable for frequency ω_{ij} representing the fraction of respondents judging stimulus *i* as stronger than *j*, while *d*’s are the positions of subsequent stimuli to be estimated.

In this manner, one obtains the actual position (mean values of normal distributions) of the stimuli and can determine the actual probabilities of judging the stimulus with the greater expected value as greater than the one with the smaller expected value. However, the normal distribution has long tails, so these probabilities will not be one. Still, they will depend on the distance between stimuli, given in units of standard deviation—the greater the distance, the higher the probability that a randomly chosen respondent will judge the stimulus with the more excellent expected value as being more outstanding.

2. Testing the Thurstone Model

A test proposed by Mosteller (1951) to assess the Thurstone model as a whole is based on the following quantity:

$$\chi^2 = \frac{\sum_{ij}(p_{ij}-\omega_{ij})^2}{\sigma_{ij}^2}, \tag{2}$$

where ω_{ij} is the observed frequency and p_{ij} is the probability based on the estimated spacing between stimuli.

According to Mosteller (1951), the quantity $\theta = \mathrm{ArcSin}\sqrt{\omega}$ has a normal distribution with variance $\sigma^2 = \frac{821}{n}$ and is nearly independent of the true p (if θ is measured in degrees). One can switch to the random variable:

$$\chi^2 = \sum_{i<j}\frac{(\theta_{ij}-\vartheta_{ij})^2}{821/n}, \tag{3}$$

with $\theta_{ij} = \mathrm{ArcSin}\sqrt{\omega_{ij}}$ and $\vartheta_{ij} = \mathrm{ArcSin}\sqrt{p_{ij}}$,

which is known to be distributed according to χ^2 with $\frac{(k-1)(k-2)}{2}$ degrees of freedom.

Thus, by comparing empirical frequencies with probabilities calculated under the assumption that the null hypothesis – the Thurstone model – holds, one obtains the value of the test statistics, which must be smaller than the critical value to uphold the null hypothesis.

Alternatively, one may test the individual distance between stimuli, not the model as a whole. Suppose the distance between two stimuli, A and B ($A < B$), measured in units of the standard deviation of the difference between these two normal random variables, is equal to d. In that case, then the probability that a randomly chosen respondent will judge B as greater than A is equal to $p = F(d)$, where F is the cumulative distribution function of a standard normal random variable.

If n independent respondents are surveyed, the probability that some (m) of them will judge *B* as greater than *A* can be described by binomial distribution $B(n,p)$ or, if n is large enough, it can be approximated by normal distribution $N\left(np, \sqrt{p(1-p)}\right)$. What is the probability that the empirical frequency will deviate from the real probability p more than δ? Clearly:

$$P(|\omega - p| > \delta) = 1 - P(p - \delta < \omega < p + \delta) = 2\left[1 - F\left(\frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}\right)\right], \tag{4}$$

with $\omega = m/n$.

If one were to test a few independent distances between stimuli simultaneously, the probabilities would multiply.

3. The Necessary Pattern of Real Probabilities Within the Thurstone Model

Let us take a moment to consider the pattern of real probabilities, assuming that the Thurstone model holds true. One of the advantages of real probabilities is the ability to arrange the matrix of these values systematically. For example, the probabilities above the diagonal are greater or equal to 0.5, increasing as one moves away from the diagonal in columns and rows (see Fig. 1).

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 22(28)

	X	*Y*	*Z*	…	*V*
X	—	> 0.5	> 0.5	> 0.5	> 0.5
Y		—	> 0.5	> 0.5	> 0.5
Z			—	> 0.5	> 0.5
…				—	> 0.5
V					—

Figure 1. Schematic pattern of probabilities for the Thurstone model

Source: own elaboration.

This pattern arises from rearranging the stimuli in ascending order – the greater the separation between stimuli (in both rows and columns), the greater the probability of correctly determining which one is greater. It is important to note that this also guarantees that the order between any two stimuli within the sequence obtained by the Thurstone method remains unaffected by the subset of stimuli chosen from the entire set for comparison. In other words, ranking any two stimuli remains independent of irrelevant alternatives [3].

Indeed, such an ordering is insufficient, as this pattern merely ensures that the stimuli are arranged increasingly rather than circularly. Moreover, it does not guarantee a linear order among them. Please refer to Fig. 2 for an illustration.

	S_1	S_2	S_3
S_1	-------	$\omega_{21} = 0.6$	$\omega_{31} = 0.9$
S_2	$1 - \omega_{21}$	------	$\omega_{32} = 0.7$
S_3	$1 - \omega_{31}$	$1 - \omega_{32}$	------

	S_1	S_2	S_3
S_1	-------	$F^{-1}(\omega_{21}) = 0.253$	$F^{-1}(\omega_{31}) = 1.282$
S_2	$1 - \omega_{21}$	------	$F^{-1}(\omega_{32}) = 0.524$
S_3	$1 - \omega_{31}$	$1 - \omega_{32}$	------

Figure 2. An example of an increasing but nonlinear order

Source: own elaboration.

While Fig. 2 displays a pattern of increasing frequencies ($\omega_{31} > \omega_{21}$ and $\omega_{31} > \omega_{32}$), it does not adhere to the condition $F^{-1}(\omega_{31}) = F^{-1}(\omega_{32}) + F^{-1}(\omega_{21})$ (as $0.253 + 0.524 \neq 1.282$), which would be expected if the observed data were linear. However, such discrepancies are not immediately apparent and require mathematical computation to detect. In contrast, Fig. 3 illustrates a pattern where the discrepancy is instantly noticeable: the distance between 3 and 1 is smaller than the sum of the distances between 3 and 2, and 2 and 1. Since both $F^{-1}(\omega_{21})$ and $F^{-1}(\omega_{32})$ alone are greater than $F^{-1}(\omega_{31})$, the former two can not sum up to the latter.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 22(28)

	S_1	S_2	S_3
S_1	-------	$\omega_{21} = 0.8$	$\omega_{31} = 0.6$
S_2	$1 - \omega_{21}$	------	$\omega_{32} = 0.9$
S_3	$1 - \omega_{31}$	$1 - \omega_{32}$	------

	S_1	S_2	S_3
S_1	-------	$F^{-1}(\omega_{21}) = 0.842$	$F^{-1}(\omega_{31}) = 0.253$
S_2	$1 - \omega_{21}$	------	$F^{-1}(\omega_{32}) = 1.282$
S_3	$1 - \omega_{31}$	$1 - \omega_{32}$	------

Figure 3. An example of nonlinear and non-increasing order

Source: own elaboration.

An even more pronounced violation of the linear Thurstone model is circularity. When we cannot arrange values above the diagonal that are all greater than or equal to 0.5, it indicates the presence of observed circularity within a group. Refer to Fig. 4 for an example.

	S_1	S_2	S_3
S_1	-------	$\omega_{21} = 0.7$	$\omega_{31} = 0.35$
S_2	$1 - \omega_{21}$	------	$\omega_{32} = 0.6$
S_3	$1 - \omega_{31}$	$1 - \omega_{32}$	------

	S_1	S_2	S_3
S_1	-------	$F^{-1}(\omega_{21}) = 0.524$	$F^{-1}(\omega_{31}) = -0.385$
S_2	$1 - \omega_{21}$	------	$F^{-1}(\omega_{32}) = 0.253$
S_3	$1 - \omega_{31}$	$1 - \omega_{32}$	------

Figure 4. An example of nonlinear and circular order

Source: own elaboration.

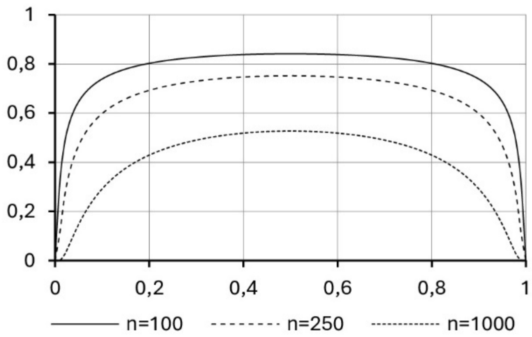
It is impossible to rearrange the data from Fig. 4, as it is impossible to obtain all frequencies above the diagonal ≥ 0.5. If we changed stimuli 1 and 3 (to obtain $\omega_{13} = 1 - \omega_{31} = 1 - 0.35 = 0.65$ above the diagonal), we would lose the other two quantities above the diagonal, changing the order and turning into values less than 0.5. This can be interpreted as circularity – the group as a whole has the following preferences: $S_2 > S_1$ (as $\omega_{21} = 0.7 < 0.5$), $S_3 > S_2$ (as $\omega_{32} = 0.6 > 0.5$), but $S_3 < S_1$ (as $\omega_{31} = 0.35 < 0.5$), which may be described as 'group circularity.'

The order depicted in Fig. 1 pertains to theoretical probabilities. By estimating the distances between stimuli based on the data, we can calculate theoretical probabilities that will undoubtedly conform to the above pattern, as they derive from the estimated linear order among stimuli. However, empirical frequencies can deviate, as they typically do not precisely match the real probabilities. This raises the question: can we estimate how much these empirical frequencies deviate from the pattern?

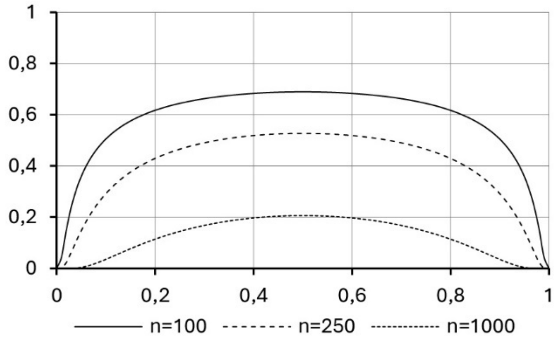
The limitation arises from the test – empirical frequencies cannot differ significantly from theoretical probabilities, as doing so would lead to rejecting the null hypothesis that the Thurstone model is true.

4. Some Estimations

Let us begin by illustrating the probabilities for a single pair of stimuli. The probability that the observed frequency will deviate from the real probability (assuming that the Thurstone model is true) is expressed by equation (4) and graphically presented in Fig. 5 for $\delta = 0.01$, $\delta = 0.02$ and p ranging from 0 to 1.



a) δ = 0.01



b) δ = 0.01

Figure 5. The probability that the observed frequency will deviate from the real probability with more than δ, for different sets of observations

Source: own elaboration.

It is evident that these probabilities decrease rapidly with both the value of deviation and the number of observations. Moreover, they would decrease even more rapidly if one were to test the differences between multiple pairs of stimuli simultaneously, as the probabilities would multiply. With a large sample size and assuming the Thurstone model holds true, the probability that observed frequencies will deviate from real probabilities is low (typically a few percentage points). For instance, considering three pairs of stimuli and $n = 250$, the probability that all three frequencies will deviate more than 2pp is on the order of $0.5^3 = 0.125$, and it decreases rapidly as the number of pairs of stimuli increases (e.g. $0.5^{k(k-1)/2}$), where k is the number of stimuli and $k(k-1)/2$ is the number of pairs.

Furthermore, one would anticipate that with increasing observations, there would be a greater alignment of empirical frequencies with the pattern imposed by the Thurstone model. We will now shift our focus to a test of the entire Thurstone model.

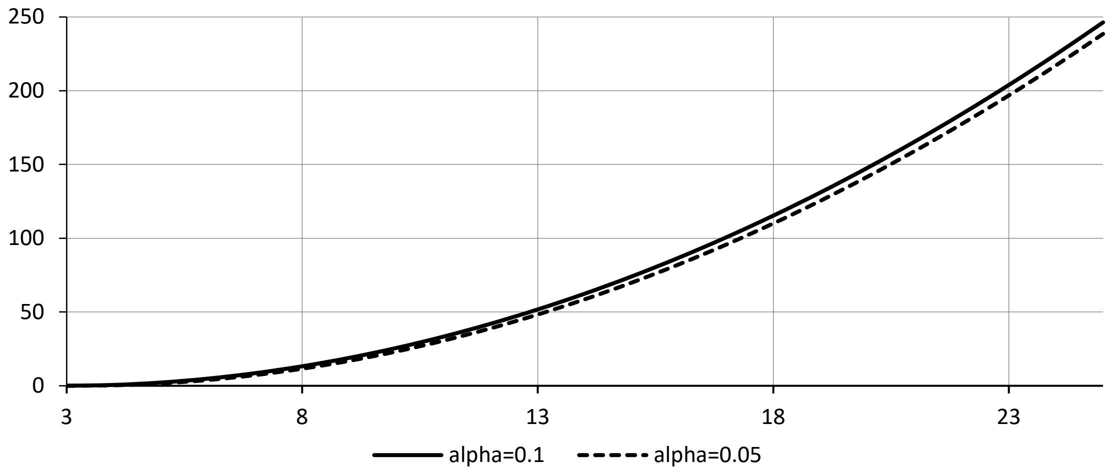


Figure 6. Critical values for *k* ranging from 3 to 25 for $\alpha = 0.1; 0.05$

Source: own elaboration.

Indeed, such an ordering is insufficient, as this pattern merely ensures that the stimuli are arranged increasingly rather than circularly. Moreover, it does not guarantee a linear order among them. Please refer to Fig. 2 for an illustration.

As both of those curves can be perfectly fitted by a polynomial of the second order:

$$y = 0.5015k^2 - 2.8358k + 3.8820, R^2 = 1 \text{ for } \alpha = 0.1$$

$$y = 0.5010k^2 - 3.1836k + 5.0253, R^2 = 1 \text{ for } \alpha = 0.05,$$

it turns out that for greater numbers of stimuli, the average squared deviation of empirical values from the theoretical probabilities should not exceed $\frac{1}{4n}$ in order not to reject the null hypothesis that the Thurstone model is true:

$$\overline{\delta^2} < \frac{1}{4n}. \tag{8}$$

This means that for $n = 100$ the value is 0.0025, and for $n = 250$ the average squared deviation is of the order of 0.001. If deviations were equal, a single deviation of the orders would yield 0.05 and 0.03, respectively.

More precise numerical values for $\alpha = 0.05$ and $\alpha = 0.1$ for $n = 250$ are presented in Fig. 7.

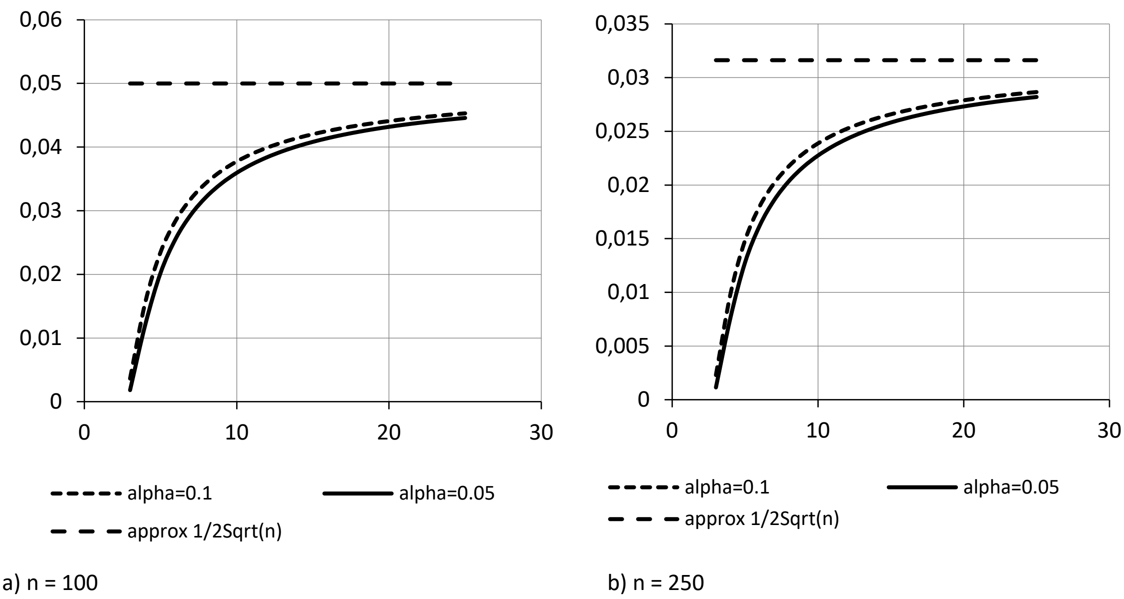


Figure 7. Average squared deviation for *k* ranging from 3 to 25 for *n* = 100 and *n* = 250, compared with the highest estimation

Source: own elaboration.

The highest limit reaches a very low value of $\overline{\delta^2} = 0.01$ for $n \approx 1500$, which is the number of respondents available for the empirical studies. Thus, the allowed average deviations from the necessary pattern are quite low.

As $4n\sum\delta^2 = \chi_c^2$, for small values of k, $k = 3,4,5$, we can also estimate the maximum deviation, assuming that $\sum\delta^2$ consists of one single deviation, which would thus be limited to the values tabularised in Tab. 1.

Table 1. Critical values and subsequent maximum values of δ

k	α	χ_c^2	$n = 100$	$n = 250$	$n = 1000$
3	0.05	0.0039	0.0031	0.0020	0.0010
	0.10	0.0158	0.0063	0.0040	0.0020
4	0.05	0.3518	0.0297	0.0188	0.0094
	0.10	0.5844	0.0382	0.0242	0.0121
5	0.05	1.6354	0.0639	0.0404	0.0202
	0.10	2.2041	0.0742	0.0469	0.0235

Source: own elaboration.

For larger values of *k*, we can employ the approximation $\delta_{\max} = \frac{k}{\sqrt{8n}}$; however, this may be less meaningful, as comparing many pairs makes it less likely that all deviations will be limited to only one of these pairs. It is preferable to state that $\sum \delta^2 \approx \frac{k^2}{8n}$.

Consequently, we can infer the following:

1. Some characteristics of the ordering of real probabilities can be easily discerned without any calculations.
2. The permissible deviations of the observed frequencies from the real probabilities are relatively small (for the number of observations typically encountered in survey research) and can be readily estimated.
3. What can be deduced from points 1) and 2) is that if the observed data deviate too significantly from the required order (which can be quantified), there is little chance that they will pass the test after fitting to the Thurstone model.

5. Examples

Let us apply the above observations to some examples. Some were created for the sake of this paper, and some were taken from previously published papers.

Example 1

The data presented in Fig. 3 violate the necessary order of increasing values both in rows and columns. The slightest possible intervention that could rectify this inconsistency is increasing ω_{31} to 0.9. This would create a deviation of 0.3. However, from the estimated orders of allowed deviations mentioned earlier, it is evident that this data will not fit the Thurstone model with sufficient accuracy to pass the test. The estimated value of the maximum allowed deviation for $k = 3$ is much smaller than 0.01 regardless of the number of observations (this holds true even for a very small number of observations, such as 10) and the level of significance. It is worth noting that our measurement of the deviation as 0.3 is only an estimation. In reality, the deviations will be distributed among three values. After fitting to the Thurstone model, it's revealed that the real $\sqrt{\sum \delta^2}$ is equal to 0.3371.

Example 2

The data presented in Fig. 4 also violates the necessary order of increasing values both in rows and columns. The slightest possible intervention that would rectify this inconsistency is increasing ω_{31} to 0.7, resulting in a deviation of 0.35. Again, the estimated value of the maximum allowed deviation for $k = 3$ is much smaller than 0.01 regardless of the number of observations and the level of significance. After fitting to the Thurstone model, it's discovered that the real $\sqrt{\sum \delta^2}$ is equal to 0.2599.

It's noteworthy that even if we focused solely on ω_{31} being less than 0.5, the minimum intervention to rectify it would be to increase it by 0.15, which still exceeds the allowed maximum deviation.

Example 3

Let us investigate the original example used by Thurstone himself, as presented in his seminal paper (1929). The empirical frequencies were based on the answers of 266 respondents. According to our estimation, the maximum allowed $\sum \delta^2$, that would not reject the null hypothesis is $\frac{19^2}{8 \cdot 266} = 0.17$. In Appendix A, we have presented some rough estimations of violations of the necessary order of the probabilities, and the sum of their squares is equal to 0.25, which is almost 1.5 times more than that 'allowed.' It is essential to note that these are very conservative estimations, as the necessary order is evidently insufficient for the linear Thurstone order to hold. Furthermore, the value of the empirical test statistics estimated with $\sum \delta^2$ is only a lower bound of the real value, as we do not take into account the higher powers of deviations. Indeed, when we run the Thurstone model on the given data, the actual $\sum \delta^2$ is equal to 0.44, resulting in the value $\tilde{\chi}_e^2 = 463$ (with a critical value equal to 131 for $\alpha = 0.1$). The exact value of test statistics, without omitting higher-order terms, is equal to $\chi_e^2 = 742$, indicating that the hypothesis of the Thurstone linear model must be rejected. However, one does not need to perform all those calculations, as the same conclusion can be reached from a very rough assessment of the proper ordering of the empirical data. While the estimations presented in Appendix A are more detailed for presentation purposes, with experience, one can quickly suspect that the data may not fit well to the linear Thurstone model in its simplest version.

Drawing from this experience, we evaluated a few more examples found in the literature and confirmed our expectations with rigorous calculations. These examples reinforce our intuition: if the data is based on a sufficiently high number of observations, for $n > 100$, the order among them should be very pronounced and of high quality to avoid rejecting the null hypothesis that the model is true after fitting to the Thurstone model.

Thurstone (1929): $n = 266$, critical value for $\alpha = 0.1$ $\chi_c^2 = 131$, empirical value $\chi_e^2 = 742$, null hypothesis rejected;

Kwan et al. (2000): $n = 200$, critical value for $\alpha = 0.1$ $\chi_c^2 = 74.2$, empirical value $\chi_e^2 = 1321.82$, null hypothesis rejected;

Krabbe (2008): $n = 212$, critical value for $\alpha = 0.1$ $\chi_c^2 = 131.0$, empirical value $\chi_e^2 = 1711.9$, null hypothesis rejected;

Dębicka et al. (2022): $n = 219$, critical value for $\alpha = 0.1$ $\chi_c^2 = 25.64$, empirical value $\chi_e^2 = 226.93$, null hypothesis rejected;

Mosteller (1951): $n = 22$, critical value for $\alpha = 0.1$ $\chi_c^2 = 13.24$, empirical value $\chi_e^2 = 14.76$, null hypothesis nearly accepted (it would be accepted at $\alpha = 0.17$).

The only example in which the null hypothesis was not rejected was for small sample size, $n = 22$. These examples support the assertion that with survey studies (larger sample sizes), obtaining empirical frequencies that are so neatly ordered is highly improbable that the null hypothesis will not be rejected after fitting into the Thurstone model. Conversely, if the frequencies were indeed ordered, the model would not need to be applied, as one could deduce the distances between the stimuli using only one row/column of the data with outstanding accuracy.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 22(28)

6. Discussion and Conclusions

From the estimations presented in the previous section, it becomes apparent that in many situations, especially when the data is based on a large number of observations, it quickly becomes evident that such data cannot be fitted to the Thurstone model with sufficient accuracy to avoid rejecting the hypothesis of the Thurstone model. In such cases, what could be the solution?

There are three possible pathways to consider. The first is to conclude outright that the Thurstone model is unsuitable and choose another tool to analyse the given data. Another potential solution is to allow for more flexibility in the model, notably by permitting different standard deviations.

The possibility of different standard deviations has been discussed since the inception of the Thurstone model. Initially, it was only possible to make estimations based on particular assumptions about which standard deviation to allow to be different from the rest. However, with the computational power available today, it is feasible to numerically minimise the differences between theoretical and empirical frequencies (as well as between theoretical and empirical spacings) and determine which standard deviations best fit the Thurstone model. While this approach may yield much lower values of χ_e^2 The status of this process is not entirely clear compared to those obtained by the standard Thurstone procedure. Is it still the Thurstone model, which was based on specific theoretical considerations about the nature of the compared stimuli? It is essential to note that if additional degrees of freedom are introduced, such as allowing the distance between each pair of stimuli to vary independently, it may result in a 100% fit of the 'model' (if it can still be named this way). However, outcomes would be ambiguous, with none highlighted as correct. Therefore, while the computational capabilities of computers are undoubtedly advantageous, one must be cautious not to overuse them.

The third possible pathway to addressing the problem of poor fit of the data to the Thurstone model is not treating it as a model that can be true or false but rather as a procedure used to impose linear order on data that are not necessarily linear by nature. Linear order acquisition, such as during elections or voting for civil projects, is inherently challenging despite the potential for inconsistencies and circularity in both individual and grouped preferences. Each system of obtaining such linear ordering may not be foolproof, and ongoing discussions persist regarding the best voting system, with different countries adopting different solutions.

If we consider the Thurstone method as one possible procedure for ordering objects, it can be compared with other methods, such as the average rank method. The average rank method is the simplest and most common, but it has limitations. It is oversimplified and does not account for possible differences in spacings between objects ranked by a single respondent as 1 and 2, 2 and 3, and so on. However, it could be justified if we impose certain assumptions on the distribution of spacings, even though these assumptions could be questioned. Additionally, the Thurstone assumption itself is also questionable, as in most situations, the data do not pass the test. The main difference between the Thurstone method and the average rank method lies in linearity. In the average rank method, the difference between 0.9 and 0.8 and between 0.3 and 0.2 (in the percentage of respondents who preferred one stimulus over another) refers to the same spacing. However, in the Thurstone model, the former has much greater dominance than the latter. This discrepancy arises from the non-linearity of normal distributions and poses challenges, especially in cases of 0 or 1 frequency, which can lead to infinite spacings in the standard Thurstone procedure. Although this particular problem can be addressed with ad hoc solutions, such as omitting these inputs from the minimised expressions, it prompts the question: what are the real advantages of the Thurstone procedure over the simple average rank method, especially in cases where the null hypothesis is rejected?

References

Baltrunas, L., Makcinskas, T., & Ricci, F. (2010). Group Recommendations with Rank Aggregation and Collaborative Filtering. In *RecSys '10: Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 119-126). ACM.

De Borda, J. (1781). Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, *102*, 657-665.

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

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank Aggregation Methods for the Web. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 613-622). ACM.

Krabbe, P. F. (2008). Thurstone Scaling as a Measurement Method to Quantify Subjective Health Outcomes. *Medical Care*, *46*(4), 357-65. https://doi.org/10.1097/MLR.0b013e31815ceca9

Kwan, Y. K., Ip, W. C., & Kwan, P. (2000). A Crime with Thurstone's Scaling of Crime Severity. *Journal of Criminal Justice*, *28*(3), 237-244.

Lin, S. (2010). Rank Aggregation Methods. *WIREs Computational Statistics*, *2*(5), 555-570. https://doi.org/10.1002/wics.111

Maydeu-Olivares, A., & Brown, A. (2010). Item Response Modelling of Paired Comparison and Ranking Data. *Multivariate Behavioral Research*, *45*(6), 935-974. https://doi.org/10.1080/00273171.2010.531231

Maydeu-Olivares, A., & Böckenholt, U. (2008). *Modelling Subjective Health Outcomes. Top 10 Reasons to Use Thurstone's Method*. Lippincott Williams & Wilkins.

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*, 207-218. https://psycnet.apa.org/doi/10.1007/BF02289116

Thurstone, L. L. (1927a). Psychophysical Analysis. *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. *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*. 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://doi.org/10.1080/01621459.1957.10501401

Model Thurstone'a: analiza danych ankietowych i ranking bodźców – model czy procedura?

Streszczenie: Niniejsza praca bada przydatność modelu Thurstone'a w badaniach preferencji. Adekwatność modelu Thurstone'a szeroko wykorzystywanego do ustalania liniowych porządków między bodźcami została poddana empirycznej walidacji. W badaniu zaproponowano nowatorski test walidacji empirycznej w celu oceny stosowalności modelu, w szczególności w odniesieniu do dużych prób. Wyniki sugerują ograniczenia w stosowaniu modelu Thurstone'a na dużych próbach. W związku z tym artykuł zawiera spostrzeżenia dla badaczy wykorzystujących ten model w celu uszeregowania bodźców i wytyczne dotyczące weryfikacji jego przydatności. Ponadto omówiono alternatywne sposoby radzenia sobie ze słabym dopasowaniem modelu, w tym uwzględnianie różnych odchyleń standardowych i traktowanie metody Thurstone'a nie jako ostatecznego modelu, ale jako procedury porządkowania obiektów. Dodatkowo badanie porównuje metodę Thurstone'a z metodą średniej rangi, podkreślając zalety i wyzwania każdego podejścia. Badanie to przyczynia się do zrozumienia złożoności badań preferencji i zapewnia cenny wgląd w udoskonalanie podejść metodologicznych w tej dziedzinie.

Słowa kluczowe: model Thurstone'a, badania preferencji, walidacja empiryczna, ranking bodźców

Appendix

Best ordering of frequencies from (Thurstone, 1929). In bold – deviations in columns underlined, and in italics – deviations in rows

	p19	p16	p5	p18	p13	p12	p4	p6	p9	p7	p8	p14	p3	p2	p11	p1	p17	p10	p15
p19		*0,939*	0,884	*0,963*	0,933	0,947	*0,928*	0,973	*0,965*	*0,958*	*0,951*	0,985	0,981	*0,966*	*0,974*	*0,955*	*0,977*	0,989	0,985
p16			***0,473***	0,525	***0,47***	0,732	*0,656*	0,779	0,805	0,801	0,859	*0,796*	0,86	0,857	***0,902***	**0,857**	0,875	0,973	0,981
p5				**0,576**	*0,506*	0,678	*0,621*	*0,764*	*0,754*	*0,745*	**0,738**	**0,728**	0,864	*0,828*	*0,924*	0,872	0,871	**0,955**	0,985
p18					**0,544**	*0,635*	**0,654**	0,716	0,74	**0,785**	*0,749*	*0,778*	**0,83**	*0,796*	*0,914*	0,826	0,879	0,974	0,977
p13						**0,652**	*0,615*	0,678	0,68	0,716	0,752	*0,702*	*0,855*	**0,818**	*0,894*	0,809	**0,886**	0,966	**0,981**
p12							0,615	*0,667*	0,657	0,697	0,695	*0,648*	0,785	0,793	0,83	***0,778***	0,848	0,97	0,97
p4								0,515	**0,534**	0,556	***0,485***	0,587	0,74	*0,757*	**0,743**	0,789	0,785	*0,97*	**0,947**
p6									0,58	**0,593**	0,605	*0,478*	**0,774**	***0,719***	***0,856***	***0,762***	*0,769*	**0,981**	0,981
p9										0,512	0,65	*0,534*	*0,746*	0,726	0,819	*0,788*	**0,82**	0,951	0,966
p7											*0,54*	0,533	0,679	0,715	*0,804*	0,756	**0,756**	**0,947**	0,963
p8												***0,474***	0,652	**0,747**	0,752	0,755	0,774	0,958	**0,977**
p14													0,651	**0,755**	*0,712*	*0,744*	0,767	0,921	0,951
p3														*0,585*	0,563	0,662	0,716	0,917	0,944
p2															***0,365***	**0,677**	*0,589*	0,863	0,925
p11																**0,682**	***0,595***	**0,917**	*0,902*
p1																	**0,419**	**0,76**	0,822
p17																		0,819	**0,924**
p10																			**0,441**

Estimated deviations in columns (with the accuracy up to centesimal parts)

		0,03	0,05	0,03	0,02	0,03		0,05	0,04	0,01	0,05	0,02	0,02	0,02	0,02	0,01	0,02	0,01
				0,04					0,04	0,12	0,03	0,03	0,01	0,09	0,01	0,05	0,01	0,02
													0,03	0,02	0,02	0,02	0,01	0,01
													0,04	0,14	0,01	0,01	0,05	0,1
															0,01	0,08	0,05	0,06

Estimated deviations in rows (with the accuracy up to centesimal parts)

		0,03	0,05	0,03	0,02	0,03		0,05	0,04	0,01	0,05	0,02	0,02	0,02	0,02	0,01	0,02	0,01
				0,04					0,04	0,12	0,03	0,03	0,01	0,09	0,01	0,05	0,01	0,02
													0,03	0,02	0,02	0,02	0,01	0,01
													0,04	0,14	0,01	0,01	0,05	0,1
															0,01	0,08	0,05	0,06