www.ees.uni.opole.pl
ISSN paper version 1642-2597
ISSN electronic version 2081-8319
Economic and Environmental Studies
Vol. 18, No 2 (46/2018), 653-664, June 2018

Ratings or pairwise comparisons? An experimental study on scale usability

Malgorzata KARPINSKA-KRAKOWIAK
University of Lodz, Poland

Abstract: A series of experiments was run in order to evaluate the usability of two different measurement approaches: ratings and ranks (pairwise comparisons). Respondents were asked to assess perceived characteristics (i.e. height and length) of different physical objects by using either a rating or a ranking scale. An artificial neural network model was built to analyse the ranks and standard statistical tests were applied to analyse the ratings. The results were then statistically compared with actual (real) characteristics of objects (i.e. their real height and length). Both systems for measuring values were found equally valid in projecting reality. Such findings offer some methodological and epistemological insights, as they provide information on the measurement power of each scale in terms of approximating real-life phenomena.
Keywords: rating scale, rank scale, pairwise comparisons, scale usability

JEL codes: *C18, C81, C89*

DOI: https://doi.org/10.25167/ees.2018.46.11

1. Introduction

Rating scales are very popular in social research. They are frequently used in various experiments and surveys in order to capture diverse phenomena e.g. individual opinions, cognitions or affective states. However common and convenient they are, ratings produce certain systematic biases and subjectivity errors, which may considerably affect final results of any research project. An alternative method of scaling is based on pairwise comparisons (pairwise preferences), where respondents are required to compare two objects and rank them. Theoretically, this approach reduces subjectivity of answers and may thus improve generalizability of data. Little

Correspondence Address: Malgorzata Karpinska-Krakowiak, Department of International Marketing and Retailing, Faculty of International and Political Studies, University of Lodz, Narutowicza 59a, 90-131 Lodz. E-mail: mkarpinska@uni.lodz.pl

has been done so far, however, to empirically evaluate this supposition and to compare ranks with other types of scales.

The general idea behind this study is to addresses the question on how much one can rely on ratings versus ranks and which measurement approach is more usable for empirical investigations in social sciences? The term "usability" is applied here to denote the degree to which a particular scale approximates an actual (i.e. real) phenomenon. In other words, the objective of the current project is to test experimentally whether ranks provide an equally true picture of reality as ratings and hence can be useful for scholarly investigations.

The present study provides numerous contributions to the literature. First, it empirically investigates the power of two measurement approaches in obtaining annotations as close to the true experiences of respondents as possible. Second, despite the existing scepticism among scholars towards rank-based data, the current study demonstrates that both ranks and ratings can be equally informative in social research. Third, it uses artificial neural networks to analyse ranks and therefore tests their applicability in capturing different phenomena beyond the field of computer information systems or computer-mediated communication. The findings reported here may be thus useful for scholars and practitioners who conduct their research in various areas, including economics, sociology, psychology, marketing, advertising, and many others.

2. Theoretical Background – Ratings and Ranks

A rating scale is based on an assumption that one can assign certain (numeric) value to the rated object. Its most popular formats include a Likert-scale (respondents are asked to indicate their level of dis/agreement with a specific statement) or a semantic differential (raters are required to evaluate an object or phenomenon in terms of opposing adjectives or phrases), and both are largely used to measure attitudes, perceptions or even behavioural intentions in many different fields, e.g. psychology (cf. Ruch and Proyer, 2009; Samson and Meyer, 2010), advertising (cf. Warren and McGraw, 2013; Yoon and Kim, 2014; Brown et al., 2010; Kim and Yoon, 2014; Kim et al., 2017), marketing or human-computer interaction studies (cf. Gosling et al., 2011; Yang, 2012; Wang, 2013). Despite their impressive popularity, ratings often yield certain risks and limitations, which seems to be left unnoticed by many researchers.

Ratings have been found to produce various systematic and personal biases. They very often force respondents to assess something abstract, imperceptible, confusing or very hard to quantify (e.g. the perceived quality of a product, funniness of an advertisement or personality of a brand), which may lead to subjectivity effects. In other words, ratings encourage individuals to evaluate things accordingly to very subjective – and thus incomparable – standards or points of reference. For example, Weijters et al. (2013) observed that the wording of end-point labels impacts subsequent responses (i.e. the more respondents use the end-point label in their daily language, the more inclined they are to select it in the questionnaire). As suggested by Linn and Gronlund (2000), left- or right-handedness may increase the tendency of an individual to select a specific side of the scale. Schwarz et al. (1991) and Cabooter et al. (2016) found that the numbering of response options may significantly affect the interpretation and use of the scale. Yannakakis and Hallam (2011) demonstrated that ratings (compared to ranks) lead to the higher recency bias (a type of order effect: when an individual is asked to assess several objects, the last one is rated higher).

Ratings are often wrongly regarded by scholars as an interval scale, i.e. the distance between subsequent response categories is presumed equal (Jamieson, 2004; Knapp, 1990). For example, while asking respondents to indicate their agreement with a statement "Advertisement X is funny", a researcher tends to assume that there exists a constant interval between "strongly agree", "agree", "disagree" and "strongly disagree". However, if there is no detailed instruction on how to define "funniness", every person that responds to this question is likely to differently assign a zero point and will probably establish different units of assessment. This may lead to personal biases and may generate some problems with further statistical analysis (see Ovadia (2004), Jamieson (2004) and Knapp (1990) for relevant examples).

The idea of rankings is to place investigated phenomena in a certain order. In the simplest form, a rank-based question requires respondents to compare two objects (e.g. to tell which one of two advertisements is funnier, better or more entertaining). Ranks may be therefore viewed as less informative than ratings, as they only provide data on ordinal relations. Neither are they unhampered by subjectivity or memory effects, as respondents may be equally inattentive or apply highly individualistic interpretations to both ranks and ratings. As reported by Wänke and Schwarz (1992), the results of ranks may be strongly influenced by the direction of comparisons and wording of the question. Additionally, rankings have been criticized for their ipsative nature, i.e. they sometimes force individuals to make choices or comparisons between things that seem

incomparable, which may thus generate individual trade-offs or unrealistic compromises (Ovadia, 2004; Dhar and Simonson, 2003).

There is an on-going debate on which measurement approach offers fewer limitations and more benefits in social research (Ovadia, 2004; Yang and Chen, 2011), but little has been done to compare these scaling systems empirically. There exists a limited number of studies that provide some insights in this field and they do not report conclusive results. For example, Ovadia (2004) demonstrates that either scale can produce equally incomplete and valid information. In their comparative survey on gaming, Yannakakis and Hallam (2011) found that there exists a varying degree of consistency between rank- and rating-based responses (i.e. correlation coefficients between ratings and preferences ranged from 0.65 to 0.92). Given such impasse, the current study is therefore designed to address the following research questions:

RQ1: Is there a difference between rank- and rating-based responses in measuring individual perceptions and judgments?

RQ2: Which measurement approach demonstrates better usability (i.e. better approximates reality)?

3. Research Design

In order to compare usability of ranks and ratings, an online experimental study with a between-subjects design was run for two types of objects. The idea was to measure the perceptions of such objects that exist in a physical world and can be evaluated both objectively (with a non-human measurement instrument) and subjectively (through a self-report delivered by respondents). This would allow comparing the real value (characteristic) of an object with perceptions captured by rank- versus rating-based questions, and to test which scaling approach better approximates reality.

Stimuli

Two different sets of physical objects were chosen for this study: trees and toothpicks. Trees are typically defined by their height (i.e. the distance between the ground and top end of branches or leaves), which can be assessed with a common metric system. Importantly, the height of a tree may be inferred from various characteristics, e.g. the size of the trunk or the number of branches and their length. Toothpicks, on the contrary, are very small wooden sticks and their simple form does not convey any additional information about their length. They, therefore, served as a control group of objects in order to test for generalizability of the findings.

The research team took photos of 5 trees that varied in their height and 5 toothpicks that varied in their length. The attempt was made to keep all the pictures comparable, e.g. all the photos were taken from the same angle, distance and with similar lighting (see Appendix 1). Each photograph was randomly numbered.

Data Collection

481 students were recruited and 465 of them completed this study. They were all enrolled in humanities, social sciences and/or marketing studies (their ages ranged from 18 to 29, M = 22.15; 72% women). They participated voluntarily in this project and they were informed about our confidentiality policy and experimental procedure. All the data were gathered and analysed anonymously; no personal information (except for age and gender) was collected. As in any other highly controlled online experiment (see e.g. Brown et al. 2010; Eisend et al., 2014; Rajabi et al. 2015) participants were asked to enter a specially designed website that contained stimuli, questions and relevant filler tasks. The research team included several attention checks and controlled for the response time as an additional quality measure (as suggested and reviewed by Guens and Pelsmacker (2017)).

The subjects were randomly assigned to one out of 15 groups representing two different conditions (each group comprised 30 up to 32 individuals). In the rating condition they were asked to take a look at a picture of a tree or a toothpick and to answer a rating-based, 5-point question (i.e. “In your opinion, how tall this tree is?”, where 0=very short, 4=very tall; “In your opinion, how long this toothpick is”, where 0=very short, 4=very long; response options were not numbered in the questionnaire). In the ranking condition respondents were presented with a pair of objects

(two trees or two toothpicks) and required to rank them. Rank-based questions were built with reference to guidelines suggested by Yannakakis and his colleagues (Yannakakis and Martinez, 2015; Martinez et al., 2014; Yannakakis and Hallam, 2011): "In your opinion, is X or Y tree taller?" and the possible answers included 3 forced-choice options (i.e. "X is taller", "Y is taller", "Both are equally tall").

4. Results

The data collection procedure resulted in 3 distinctive datasets for each investigated object, i.e.: (1) data collected by means of a semantic differential scale; (2) rank-based data; (3) data describing real height and length values. Despite certain controversy around the statistical approach to ratings (Jamieson, 2004; Yannakakis and Martinez, 2015), the author decided to follow a dominant practice in social research and calculate mean, median and standard deviation values across subjects in the rating-based dataset.

In order to analyse rankings, the preference learning toolbox (PLT) was used (Farrugia et al., 2015). PLT derives from machine learning field and it allows building an artificial neural network model that predicts the effects of ranks. In the present study, a 2-layer perceptron was built for each type of object. In each case, a 5-2-1 architecture was applied (i.e.: 5 neurons constituted an input layer; there were 2 neurons in a single hidden layer and a neuron on the output). A sigmoid function was used as an activation for hidden and output layers, thus it produced a numerical value (between 0 and 1) for each investigated object (see Table 1). After a few training sessions the models with the highest accuracy rates were chosen (92% and 97%).

Table 1. Ranks, ratings and real characteristics of investigated objects

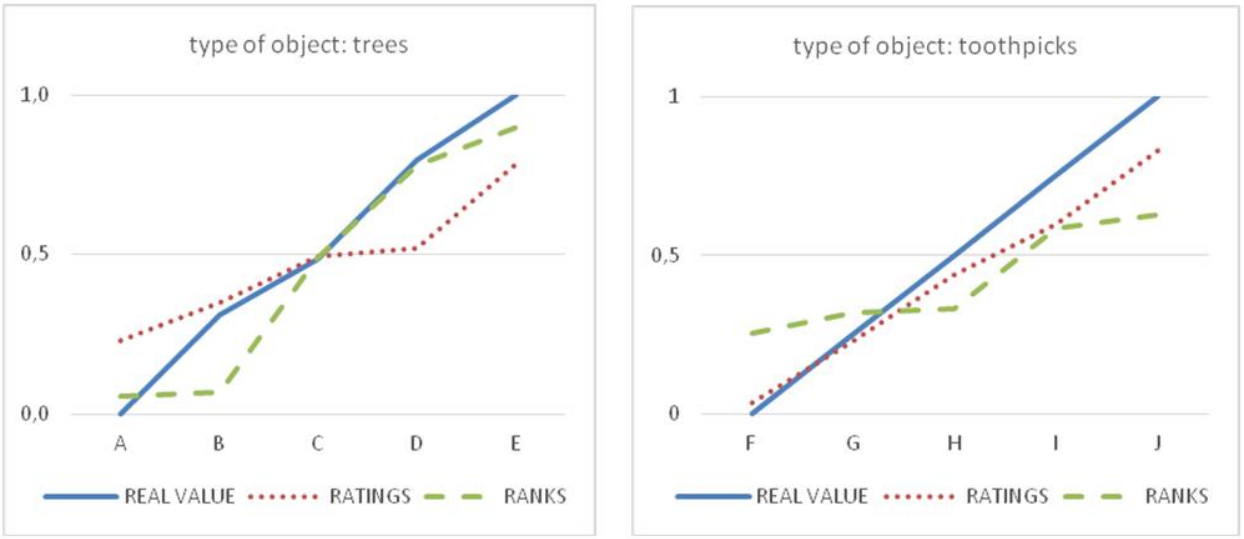
TREES	**REAL VALUE** (height in meters)	**RATING** (mean [median] values)	**RANKING** (PLT results)
A	2.59	0.92 [0] (SD=1.05)	0.05
B	3.75	1.39 [1] (SD=0.63)	0.07
C	4.41	1.97 [2] (SD=0.63)	0.49
D	5.57	2.08 [2] (SD=0.63)	0.78
E	6.33	3.13 [3] (SD=0.77)	0.90
TOOTHPICKS	**REAL VALUE** (length in centimeters)	**RATING** (mean [median] values)	**RANKING** (PLT results)
F	1	0.14 [0] (SD=0.47)	0.25

G	2	0.92 [1] (SD=0.64)	0.32
H	3	1.75 [2] (SD=0.58)	0.33
I	4	2.39 [2] (SD=0.73)	0.58
J	5	3.31 [3] (SD=0.68)	0.63

Source: Author's own elaboration

The next step was to transform the results (i.e. numerical values assigned to each investigated object by means of either ratings or ranks – see Table 1) in a way to make them comparable. Each value in a rating dataset was therefore divided by the scale length in order to obtain values between 0 and 1 (as in the ranking dataset). Then an independent samples Mann-Whitney test was performed to statistically compare the results of rank- versus rating-based questions. No significant differences were found between responses measured by ranks and ratings with regard to both kinds of physical objects (trees: Mann-Whitney U = 11, Z=-0.31, p=0.75; toothpicks: Mann-Whitney U = 12, Z=-0.10, p=0.91). In both cases correlation coefficients (Spearman's rho) were above 0.91 and p-values were ≤0.02. This implies that both measurement methods worked similarly and were equally usable in approximating reality (see Figure 1). Such results address RQ1 and RQ2.

Figure 1. Ranks, ratings and real characteristics of investigated objects



Source: Author's own elaboration

5. Discussion and Conclusions

The popularity of ratings in marketing and consumer research is overwhelming (see Bruner, 2009). Many experiments and surveys rely heavily on Likert or semantic differential scales. In the advertising literature, for example, it is not uncommon to ask respondents about the extent to which they perceive particular brand, product or ad as funny (e.g. Warren and McGraw, 2013; Cline et al., 2003; Yoon and Kim, 2014), pleasing (e.g. Das et al., 2015), violent (e.g. Brown et al., 2010; Kim and Yoon, 2014), truthful (e.g. Kim et al., 2017), hedonic (e.g. Voss et al., 2003), involving (e.g. Stokburger et al., 2012), etc. Given high potential subjectivity of such questions and personal bias that might stem from them, it is worth finding out how ratings actually work in comparison with alternative scaling approaches, and how well they illustrate the ground truth.

The present study contributes to the existing body of literature in various ways. First, to the best of the author's knowledge, it is one of the first empirical attempts to experimentally compare the usability of pairwise ranks and ratings. By introducing physical objects into the investigation, it was possible to inspect the distance between results given by different scales (i.e. human perceptions) and reality. The findings of the present study suggest that both ranks and ratings may be similarly consistent in predicting real-life phenomena. Even though each scale may work a little differently for various groups of objects (see Figure 1), the overall performance seems similar.

Second, the present study introduces ranks into the stream of research that has so far been dominated by rating-based investigations. The current findings demonstrate that both ranks and ratings may be equally informative and good enough in predicting the correct order of various phenomena. If ranks produce less cognitive overload (an assumption forwarded by Yang and Chen [2011]) and fewer reporting biases (Yannakakis and Hallam, 2011), they may be more effective and comfortable to use than ratings (especially in such surveys or experiments that require high attention and memory capacity from respondents).

Third, by introducing preference learning methods into current analysis the present study provides certain information about PLT validity in social research. Even though artificial neural networks have been extensively used in various fields (e.g. sales and investment estimations or gaming industry), they are largely uncommon in social research, especially in advertising, branding or marketing communications. Based on the current findings one may conclude that using PLT to predict ranks can be as easy and informative as rating-based analyses and can additionally help omit potential statistical errors (Ovadia 2004).

The current experiment is not deprived of certain limitations. It only focuses on how individuals perceive physical characteristics of physical objects and how they report their cognitions in relation to reality. It does not investigate affective judgments or preferences. It is therefore highly important to further test usability of ranks versus ratings in approximating not only various emotional states, but also more complicated and less tangible phenomena (e.g. consumer satisfaction, liking or anxiety). Future studies should also examine a wider range of circumstances under which either type of the scale provides better responses.

Literature

Brown, M.R., R.K. Bhadury, N. Pope (2010). The impact of comedic violence on viral advertising effectiveness. Journal of Advertising 39: 49-66.

Bruner, G (2009). Marketing Scales Handbook: A Compilation of Multi-Item Measures for Consumer Behavior & Advertising Research, vol. 5, GCBII Productions, Carbondale, Illinois.

Cabooter, E., B. Weijters, M. Geuens, I. Vermeir (2016). Scale Format Effects on Response Option Interpretation and Use. Journal of Business Research 69(7): 2574-2584.

Cline, T.W., M.B. Altsech, J.J. Kellaris (2003). When Does Humor Enhance or Inhibit Ad Responses? The Moderating Role of the Need for Humor. Journal of Advertising 32(3): 31-45.

Das, E., M. Galekh, C. Vonkeman (2015). Is sexy better than funny? Disentangling the persuasive effects of pleasure and arousal across sex and humour appeals. International Journal of Advertising 34(3): 406-420.

Dhar, R., I. Simonson (2003). The Effect of Forced Choice on Choice. Journal of Marketing Research 40(2): 146-160.

Eisend, M., J. Plagemann, J. Sollwedel (2014). Gender Roles and Humor in Advertising: The Occurrence of Stereotyping in Humorous and Nonhumorous Advertising and Its Consequences for Advertising Effectiveness. Journal of Advertising, 43(3), 256-273.

Farrugia, V.E., H.P. Martínez, G.N. Yannakakis (2015). The preference learning toolbox. [arXiv:1506.01709].

Geuens, M., P. De Pelsmacker (2017). Planning and Conducting Experimental Advertising Research and Questionnaire Design. Journal of Advertising, 46(1), 83-100.

Gosling, S.D., A. Augustine, S. Vazire, N. Holtzman, S. Gaddis (2011). Manifestations of Personality in Online Social Networks: Self-Reported Facebook-Related Behaviors and Observable Profile Information. Cyberpsychology, Behavior and Social Networking 14: 483-488.

Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education 38: 1217-1218.

Kim, E., S. Ratneshwar, E. Thorson (2017). Why Narrative Ads Work: An Integrated Process Explanation. Journal of Advertising 46(2): 283-296.

Kim, Y., H. Yoon (2014). What Makes People <Like> Comedic-Violence Advertisements? A Model for Predicting Attitude and Sharing Intention. Journal of Advertising Research June: 217-232.

Knapp, T. (1990). Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy. Nursing Research 39(2): 121-123.

Linn,,R., N. Gronlund (2000). Measurement and Assessment in Teaching. Upper Saddle River, NJ: Prentice-Hall

Martinez, H.P., G.N. Yannakakis, J. Hallam (2014). Don't Classify Ratings of Affect; Rank them! IEEE Transactions on Affective Computing 1-14.

Ovadia, S. (2004). Ratings and rankings: reconsidering the structure of values and their measurement. International Journal of Social Research Methodology 7(5): 403–414.

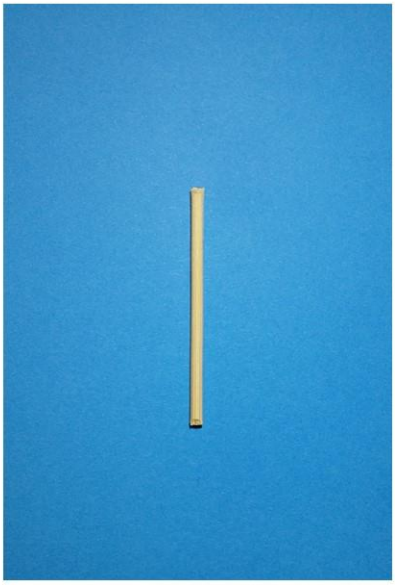
Rajabi, M., N. Dens, P. De Pelsmacker, P. Goos (2015). Consumer responses to different degrees of advertising adaptation: the moderating role of national openness to foreign markets. International Journal of Advertising, doi:10.1080/02650487.2015.1110949.

Ruch, W., R.T. Proyer. (2009). Extending the study of gelotophobia: On gelotophiles and katagelasticists. Humor – International Journal of Humor Research 22, 1/2: 183-212.

Samson, A.C., Y. Meyer (2010). Perception of aggressive humor in relation to gelotophobia, gelotophilia, and katagelasticism. Psychological Test and Assessment Modeling 52(2): 217-230.

Schwarz, N., B. Knäuper, H. Hippler, E. Noelle-Neumann, L. Clark (1991). Rating Scales: Numeric Values May Change the Meaning of Scale Labels. Public Opinion Quarterly, 55(4), 570-582.
Stokburger-Sauer, N., S. Ratneshwar, S. Sen (2012). Drivers of Consumer-Brand Identification. International Journal of Research in Marketing 29(4): 406-418.
Voss, K.E., E.R. Spangenberg, and B. Grohmann (2003). Measuring the Hedonic and Utilitarian Dimensions of Consumer Attitude. Journal of Marketing Research 40(3): 310–320.
Wang, S.S. (2013). I Share, Therefore I Am: Personality Traits, Life Satisfaction, and Facebook Check-Ins. Cyberpsychology, Behavior, and Social Networking 16: 870-877.
Wänke, M., N. Schwarz (1992). Comparative judgements: how the direction of comparison determines the answer. Conference paper; Zentrum für Umfragen, Methoden und Analysen –ZUMA, Mannheim. Available at: file:///C:/Users/User/Downloads/ssoar-1992-wanke_et_al-comparative_judgements_how_the_direction.pdf. Accessed 29 March 2018.
Warren, C., P. McGraw (2013). When humor backfires: revisiting the relationship between humorous marketing and brand attitude. Marketing Science Institute Working Paper Series 13(124): 1-40.
Weijters, B., M. Geuens, H. Baumgartner (2013). The Effect of Familiarity with the Response Category Labels on Item Response to Likert Scales. Journal of Consumer Research 40(2): 368-381.
Yang, T. (2012). The Decision Behavior of Facebook Users. Journal of Computer Information Systems 52: 50-59.
Yang, Y.H., H.H. Chen (2011). Ranking-based emotion recognition for music organization and retrieval. IEEE Transactions on Audio, Speech, and Language Processing 19: 762-774.
Yannakakis, G.N., H.P. Martinez (2015). Ratings are Overrated! Frontiers in ICT 2(13): 1-5.
Yannakakis, G.N., J. Hallam (2011). Rating vs. preference: A comparative study of self-reporting. Affective Computing and Intelligent Interaction, 6974: 437-446.
Yoon, H.J., Y. Kim (2014). The Moderating Role of Gender Identity in Responses to Comedic Violence Advertising. Journal of Advertising 43: 382-396.

Appendix 1. Sample stimuli used in the study



Skale ratingowe czy porównania parami? Badania eksperymentalne nad użytecznością skal pomiarowych

Streszczenie

Celem badania była ocena użyteczności dwóch różnych skal pomiarowych (ratingowej i rangowej, czyli tzw. porównań parami) w zakresie aproksymacji rzeczywistych zjawisk. Uczestnicy eksperymentu zostali losowo podzieleni na kilka grup badawczych, w ramach których mieli za zadanie ocenić postrzegane cechy różnych obiektów (tj. wysokość lub długość). Ocen tych dokonywali za pomocą jednej z badanych skal pomiarowych. W celu analizy wyników skali rangowej zbudowano model sztucznej sieci neuronowej, natomiast do zanalizowania rezultatów skali ratingowej zastosowano standardowe testy statystyczne. Następnie wszystkie wyniki porównywano z rzeczywistymi charakterystykami obiektów (tj. ich prawdziwą wysokością lub długością). Oba systemy pomiaru okazały się równie dobre w aproksymacji rzeczywistości. Eksperyment ten niesie pewną wartość poznawczą w kontekście metodologicznym oraz epistemologicznym, ponieważ dostarcza informacji o sile pomiarowej każdej ze skal pod względem trafności odzwierciedlania badanych zjawisk.

Słowa kluczowe: skala porządkowa, skala rangowa, użyteczność skal, porównania parami.