JPO, Vol. 5(3), 17-23

Journal of Policy Options · RESDO

Comparing the Validity of Rating Scales and Ranking Methods in Measuring Perceived Characteristics

Han Lee[a], Kim Hur[b]

Abstract

A series of experiments was conducted to evaluate the usability and effectiveness of rating scales and ranking methods. Participants were tasked with assessing perceived characteristics, specifically the height and length of various physical objects, using either a rating scale or a ranking system. To analyze the data, an artificial neural network model was developed for the rankings, while standard statistical tests were employed to evaluate the ratings. The core objective of the study was to determine how accurately these measurement systems reflect actual physical characteristics by statistically comparing the respondents' assessments to the objects' real dimensions. The findings revealed that both rating and ranking approaches are equally valid in their ability to project reality. This equivalence in validity suggests that each method is capable of approximating real-life phenomena with a similar degree of accuracy. The implications of these results are significant for both methodological and epistemological considerations. Methodologically, the study highlights the robustness of both measurement techniques in capturing perceptual data that closely aligns with actual physical properties. Epistemologically, it provides insights into the power of different scales to measure and approximate real-world phenomena, suggesting that researchers can choose between these approaches based on preference or context without sacrificing accuracy. These findings contribute to a deeper understanding of the strengths and limitations of different measurement systems, particularly in fields where precise and accurate data collection is crucial. By demonstrating the comparable validity of ratings and ranks, this research offers valuable guidance for future studies that seek to measure and interpret perceived characteristics, ensuring that the chosen method is well-suited to the specific requirements of the research context.

Keywords: Rating Scales, Ranking Methods, Measurement Validity

JEL Codes: C83, D91, C45

1. INTRODUCTION

Rating scales are widely used in social research as they provide a convenient and effective way to quantify subjective responses. They are commonly employed in experiments and surveys to capture a variety of phenomena, including individual opinions, perceptions, attitudes, and emotional states. By allowing respondents to express the intensity or degree of their feelings or thoughts on a particular topic, rating scales enable researchers to analyze complex social and psychological variables in a structured and comparable manner. This makes them an essential tool for understanding diverse aspects of human behavior and cognition in both qualitative and quantitative studies. While rating scales are commonly used and convenient, they can introduce systematic biases and subjectivity errors that may significantly impact the accuracy and reliability of research results. Respondents may interpret scale points differently or be influenced by external factors, leading to variations that do not accurately reflect their true opinions or experiences. An alternative approach to scaling is based on pairwise comparisons, also known as pairwise preferences. In this method, respondents are asked to compare two objects at a time and rank them relative to each other. This forces a direct comparison between the options, which can reduce the cognitive load associated with rating multiple items simultaneously and minimize some of the biases inherent in traditional rating scales. Pairwise comparisons provide a more objective way of capturing preferences, as respondents focus on specific distinctions between pairs, making it less likely that they will be influenced by irrelevant factors or personal interpretation of scale values.

Theoretically, the pairwise comparison approach is believed to reduce the subjectivity inherent in traditional rating scales, potentially leading to more reliable and generalizable data. By focusing respondents on direct comparisons between two options, it minimizes the ambiguity that often arises when interpreting rating scales. This method can reduce the influence of individual biases, such as scale-use tendencies or differences in interpretation of numerical values. However, despite these theoretical advantages, there has been limited empirical research to evaluate the effectiveness of pairwise comparisons in practice. Few studies have systematically compared how ranks derived from pairwise comparisons stack up against other types of scaling methods, such as Likert scales or numerical ratings. More research is needed to understand whether pairwise comparisons consistently lead to better data quality, improved reliability, and greater generalizability across different types of research settings and respondent populations. This gap in empirical evaluation suggests that while

[a] School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China
[b] School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China

promising, pairwise comparisons warrant further investigation before they can be fully endorsed as a superior alternative. The primary aim of this study is to explore the reliability and usability of ratings versus ranks in empirical social science research. The key question addressed is how much one can depend on these two measurement approaches to accurately capture real-world phenomena. In this context, "usability" refers to the extent to which a particular scale effectively approximates the actual or real nature of the phenomenon under investigation. Essentially, the study seeks to determine whether ranks, as an alternative to ratings, provide an equally valid representation of reality and can thus serve as a useful tool in academic research. By experimentally testing the effectiveness of ranks in comparison to ratings, the study aims to contribute to the ongoing discussion about which measurement approach is more appropriate for different types of social science investigations. The findings will help determine whether ranks, which theoretically reduce bias and subjectivity, offer a more accurate and reliable picture of empirical phenomena than traditional rating scales, and whether they can be more broadly adopted in scholarly research.

2. LITERATURE REVIEW

A rating scale operates on the assumption that a numeric value can be assigned to the object being evaluated. Common formats include the Likert scale, where respondents indicate their level of agreement or disagreement with a specific statement, and the semantic differential, where respondents evaluate an object or phenomenon using opposing adjectives or phrases. These formats are widely used across various fields to measure attitudes, perceptions, and even behavioral intentions. For instance, in 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, and human-computer interaction studies (Gosling et al., 2011; Yang, 2012; Wang, 2013), rating scales have become an essential tool. Despite their widespread use and popularity, ratings are not without their risks and limitations. One major issue is that they may introduce bias, such as respondents interpreting the scale points differently or being influenced by external factors, which can distort the results. Additionally, respondents may tend to choose central or extreme options based on their personalities or cultural tendencies, leading to potential inaccuracies. These risks are often overlooked, as researchers frequently rely on the convenience and familiarity of rating scales without fully considering the potential limitations. As a result, while ratings can provide useful data, they may not always yield the most accurate or reliable representation of respondents' true feelings or behaviors. This makes it important to critically assess when and how to use ratings, and to explore alternative methods, such as ranks, that might mitigate these issues.

Rating scales have been shown to introduce various systematic and personal biases, which can distort the accuracy of the data collected. One major issue is that ratings often require respondents to evaluate abstract, difficult-to-quantify concepts such as the perceived quality of a product, the funniness of an advertisement, or the personality of a brand. This can lead to subjectivity effects, where individuals rely on personal, and therefore incomparable, standards or points of reference. As a result, ratings may not reflect an objective measure but rather a personal interpretation of the concept being rated. For instance, Weijters et al. (2013) found that the wording of end-point labels on a scale can significantly influence responses, with respondents more likely to select an option if it aligns with the language they commonly use. Similarly, Linn and Gronlund (2000) suggested that a respondent's left- or right-handedness might increase the likelihood of them selecting a particular side of the scale, introducing another layer of subjectivity. Schwarz et al. (1991) and Cabooter et al. (2016) highlighted that the numbering of response options can affect how respondents interpret and use the scale, leading to potential inconsistencies in responses across different contexts. Moreover, Yannakakis and Hallam (2011) demonstrated that ratings, as compared to ranking methods, are more susceptible to recency bias—a type of order effect where respondents tend to give higher ratings to the last object they evaluate in a series. This suggests that the sequential order in which options are presented can influence the outcome, further undermining the objectivity of rating scales. These findings collectively point to the limitations of ratings, which, while convenient and widely used, may introduce biases that compromise the reliability and comparability of the data. Ratings are often mistakenly treated as interval scales by researchers, meaning that the distance between subsequent response categories is presumed to be equal (Jamieson, 2004; Knapp, 1990). For instance, when respondents are asked to rate their agreement with a statement such as "Advertisement X is funny," researchers commonly assume that the intervals between response options like "strongly agree," "agree," "disagree," and "strongly disagree" are uniform. This assumption implies that the difference between each category is the same, allowing for the use of parametric statistical methods in the analysis.

However, in reality, without a clear and standardized definition of subjective concepts like "funniness," each respondent is likely to interpret and assign their own zero point and measurement units. This lack of consistency can result in varying scales of assessment across individuals, leading to personal biases. One person's "agree" may be another's "strongly agree," and the perceived gap between "disagree" and "strongly disagree" could differ widely between respondents. This variation in interpretation presents challenges for researchers, especially when they attempt to perform statistical analysis based on the assumption of equal intervals. As pointed out by Ovadia (2004), Jamieson (2004), and Knapp (1990), treating ratings as interval data when they are actually ordinal can lead to skewed results and improper conclusions. Therefore, it is critical to recognize the limitations of rating scales and to consider whether other methods, such as ranks or more clearly defined scales, might provide a more reliable and valid approach for measuring subjective responses.

The concept of rankings involves placing phenomena or objects in a specific order based on certain criteria, often asking respondents to judge which object is preferable, better, or more fitting according to a particular dimension. In its simplest

form, a rank-based question asks respondents to compare two objects, such as determining which of two advertisements is funnier, more memorable, or more entertaining. While this method provides clear ordinal relationships between items, it is often viewed as less informative than ratings because rankings do not capture the intensity of preferences—only the relative positions of the objects are revealed. For example, if someone ranks two products as first and second, the data does not indicate whether the difference between them is marginal or substantial. This lack of nuance is one reason rankings are sometimes considered limited in comparison to rating scales, which allow respondents to indicate the degree to which they prefer or dislike something. Ratings provide more granular data, offering insights into not just which option is preferred but by how much.

Moreover, like ratings, rankings are not entirely free from subjectivity or bias. Although ranking eliminates the need for respondents to interpret abstract numerical scales, it does not fully mitigate issues such as personal biases or memory effects. Respondents may still display inattentiveness, inconsistent judgments, or unique personal interpretations when comparing items. For instance, they might apply subjective criteria, prioritizing aspects that other respondents may not find important. Research by Wänke and Schwarz (1992) highlights that rankings can be heavily influenced by the direction of the comparison and the specific wording of the questions. Subtle differences in how a comparison is framed may cause respondents to focus on different attributes of the objects, skewing the results. For instance, asking respondents to choose the "better" advertisement versus the "more entertaining" one may lead to different rankings even though both questions ostensibly ask for a comparison. Additionally, rankings face another significant critique related to their ipsative nature. This means that rankings can sometimes force respondents to make choices between objects that may not naturally lend themselves to direct comparison. For example, a respondent may be asked to rank several distinct types of products or advertisements, but in reality, these items may appeal to different needs or preferences, making them hard to compare on the same scale. As a result, respondents may make trade-offs or compromises that do not truly reflect their preferences, leading to distorted results. Ovadia (2004) and Dhar and Simonson (2003) argue that this forced comparison can lead respondents to create artificial distinctions between objects that they might otherwise view as equally favorable or unfavorable. Such forced comparisons can result in rankings that do not fully capture a respondent's actual preferences, as they might have been coerced into ranking items that seem incomparable. Furthermore, rankings require respondents to place each object in a specific order, which may be difficult when the differences between the objects are minor or when they serve different purposes. For example, ranking several brands of cars or advertisements could force a respondent to assign a relative position even when they consider multiple items to be equally favorable or when they find certain attributes unimportant for the context of comparison. This task can generate frustration or cognitive overload, especially when respondents perceive all the options as either similarly appealing or irrelevant to them. Thus, the cognitive demands placed on respondents can affect the accuracy of the rankings, especially if they are unsure how to prioritize certain criteria or how to make sense of incomparable items.

Another limitation of rankings is that they sometimes mask important insights about the actual reasons behind preferences. Unlike ratings, which allow respondents to express how strongly they feel about something, rankings are purely ordinal and do not reveal the magnitude of preference. This lack of information about preference strength means that two individuals who rank the same items in the same order could have vastly different feelings about those items. One person might have very strong preferences between options, while another might see them as virtually equal. Without knowing the degree of difference, researchers miss out on key data that could influence interpretations and subsequent decisions. Despite these drawbacks, rankings can offer some advantages, especially when the goal is to understand relative preferences without the influence of abstract numerical scales, which can introduce additional biases. Rankings force direct comparisons, which can reduce ambiguity in cases where respondents may struggle with rating scales or when precise numeric differentiation is not required. For example, in marketing or consumer research, rankings can be useful for understanding which products or advertisements stand out in competitive markets. Rankings are also less affected by extreme response tendencies, such as respondents always choosing the highest or lowest points on a rating scale, which can artificially inflate or deflate the results in ratings.

However, it is clear that rankings, like ratings, are subject to their own set of challenges. While rankings may reduce some subjectivity, they are not without their limitations, including biases introduced by question framing, the difficulty of making comparisons between incomparable objects, and the inability to capture the intensity of preferences. Additionally, the ipsative nature of rankings may result in unrealistic trade-offs, and the cognitive load required to rank items may influence the accuracy of responses. While rankings provide a useful tool for understanding relative preferences, they are not immune to the challenges faced by other measurement approaches, such as ratings. Both methods have their advantages and limitations, and the choice between them should be carefully considered based on the research context, the nature of the items being compared, and the type of insights the researcher seeks to obtain. While rankings can offer clarity in comparisons, they may fall short in capturing the full depth of respondents' preferences, and researchers must remain mindful of the biases and cognitive demands associated with this method. There is an ongoing debate about which measurement approach—ratings or rankings—offers fewer limitations and more benefits in social research (Ovadia, 2004; Yang and Chen, 2011). Despite the significance of this question, there has been relatively little empirical work comparing these two scaling systems directly, and the existing studies have not reached conclusive results. This gap in the literature leaves researchers uncertain about which method is more reliable or effective in capturing nuanced social phenomena.

For instance, Ovadia (2004) found that both rating and ranking scales can produce incomplete yet valid information, suggesting that neither method is inherently superior. This implies that both approaches have their strengths and weaknesses, and their effectiveness may depend on the context in which they are used. However, this study did not point decisively to one method being more reliable or informative than the other. In a comparative survey on gaming, Yannakakis and Hallam (2011) investigated the relationship between rank- and rating-based responses and found varying degrees of consistency between the two approaches. Their results showed that correlation coefficients between ratings and preferences ranged from 0.65 to 0.92, indicating that while the two methods were often aligned, they were not always perfectly consistent. This variation suggests that individual respondents may interpret or respond to ranks and ratings differently, depending on the task or the subject matter. The few studies that have examined the differences between these measurement approaches suggest that both have their place in social research but may produce different types of data depending on the research context. However, the lack of empirical consensus indicates a need for further comparative studies to better understand the nuances of these methods and to provide more guidance on when and how each should be used in research.

3. METHODOLOGY

To compare the usability of ranks and ratings, an online experimental study using a between-subjects design was conducted with two types of physical objects: trees and toothpicks. The aim was to measure perceptions of these objects, both in terms of objective measurements (using a non-human instrument) and subjective assessments (through self-reports from respondents). This approach allowed researchers to compare the actual physical characteristics of the objects with perceptions gathered through rank-based and rating-based questions, thereby determining which scaling method better approximates reality. The two object types were chosen to highlight different levels of perceptual complexity. Trees, which vary in height, were selected as an object that can be evaluated based on multiple visual cues, such as trunk size, number of branches, and overall appearance. These visual characteristics could lead to subjective interpretations of height, making trees a suitable object for testing how well ranks and ratings capture complex perceptions. On the other hand, toothpicks served as the control group due to their simplicity. As small, uniform sticks, toothpicks do not convey much additional information beyond their length, allowing for a cleaner test of how respondents evaluate simple objects with limited perceptual ambiguity. A total of 481 students were recruited for the study, of whom 465 completed it. These participants were all enrolled in humanities, social sciences, and marketing programs, with ages ranging from 18 to 29 (mean age = 22.15; 72% women). Participation was voluntary, and all participants were informed about the confidentiality policy and experimental procedures. No personal information beyond age and gender was collected, ensuring the anonymity of the data. The study was designed and administered via a specially constructed website that presented stimuli, questions, and filler tasks to the participants.

As with any highly controlled online experiment (e.g., Brown et al., 2010; Eisend et al., 2014; Rajabi et al., 2015), the research team took several precautions to ensure the quality of the data. Attention checks were embedded within the study to filter out inattentive responses. Additionally, response times were tracked as an extra quality control measure, following the suggestions reviewed by Guens and Pelsmacker (2017). This combination of attention checks and response time monitoring helped maintain data integrity, ensuring that participants engaged with the task and provided thoughtful responses. By using these two sets of objects—complex (trees) and simple (toothpicks)—the study aimed to generalize findings across different levels of perceptual difficulty. The experimental design allowed for a direct comparison between how well ranks and ratings approximate the actual physical characteristics of the objects, thus providing insights into the reliability and usability of these scaling methods in empirical social research. The results of this study have the potential to inform researchers about which method may be more appropriate depending on the type of object or phenomenon being evaluated, as well as the context in which the evaluation is conducted.

4. OUTCOMES

The data collection process yielded three distinct datasets for each investigated object (trees and toothpicks): (1) data collected using a semantic differential scale (ratings), (2) rank-based data, and (3) data representing the real height (for trees) and length (for toothpicks) values. Although there is ongoing debate regarding the statistical treatment of rating scales, particularly in terms of whether they should be treated as interval data (Jamieson, 2004; Yannakakis and Martinez, 2015), the study followed the common practice in social research. As such, the mean, median, and standard deviation values were calculated for the rating-based dataset, ensuring consistency with the dominant analytical methods used in similar research. For the analysis of the rank-based data, the *Preference Learning Toolbox* (PLT) was employed (Farrugia et al., 2015). PLT is derived from the field of machine learning and is particularly useful for modeling preferences and predicting outcomes based on ranked data. It enables the construction of artificial neural network models that can analyze the effects of rankings. In this study, a two-layer perceptron model was built for each type of object—trees and toothpicks—allowing the researchers to predict outcomes based on the rank-based data. For both types of objects, the same neural network architecture was used: a 5-2-1 architecture, meaning that the input layer contained five neurons, the hidden layer contained two neurons, and the output layer consisted of a single neuron. A sigmoid function was employed as the activation function for both the hidden and output layers. This function produced a numerical value between 0 and 1 for each object, which was then used to quantify the prediction results (see Table 1 for output examples).

After several training sessions, the models were optimized to achieve high accuracy rates. The final models chosen for the analysis reached an accuracy of 92% for trees and 97% for toothpicks. These high accuracy rates indicate that the artificial neural network models effectively captured the relationships within the rank-based data and were able to predict the outcomes with substantial reliability. By employing both traditional statistical methods for the rating-based data and advanced machine learning techniques for the rank-based data, the study was able to compare how well each scaling method approximates the real values of the physical objects under investigation. This comprehensive approach allowed for a deeper exploration of the usability of rankings versus ratings in social research, providing insights into the strengths and limitations of each method in capturing objective reality. The use of PLT also highlighted the potential of machine learning models to enhance the analysis of rank-based data in empirical investigations.

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

The table presents the ranks, ratings, and real values of two sets of investigated objects: trees and toothpicks. For each object, real values (heights in meters for trees and lengths in centimeters for toothpicks) are compared to their corresponding ratings and rankings. The trees (A to E) have real heights ranging from 2.59 meters (Tree A) to 6.33 meters (Tree E). The ratings, represented as mean [median] values with standard deviations (sd), reflect how these trees are perceived or rated on a scale, likely based on some characteristic like size or appeal. For instance, Tree A has a mean rating of 0.92 with a median of 0, indicating lower perception relative to others, while Tree E, with the tallest real value, has a higher mean rating of 3.13 with a median of 3. The ranking (plt results) values represent their relative standing or preference, with lower values indicating a higher rank. Tree A, despite its low rating, has a ranking value of 0.05, suggesting it is relatively highly ranked within its group, while Tree E, with the highest height, has a ranking of 0.90, reflecting a lower preference despite its physical stature.
For the toothpicks (F to J), real lengths range from 1 cm (Toothpick F) to 5 cm (Toothpick J). Ratings are provided similarly, showing the perceived value of these objects, with mean and median values reflecting the central tendency of their ratings. Toothpick F, with the shortest length, has a mean rating of 0.14 and a median of 0, indicating a low perception, while Toothpick J, with the longest length, has a mean rating of 3.31 and a median of 3. The ranking values indicate the relative standing of each toothpick, where Toothpick F has a ranking of 0.25 and Toothpick J has a ranking of 0.63, suggesting that while J is rated higher, it is ranked lower relative to some other objects in the group. Overall, the table illustrates how the real values (physical measurements) of these objects correlate with their perceived ratings and assigned ranks. It highlights potential discrepancies between physical characteristics and subjective evaluations, showing that taller or longer objects are not always rated or ranked the highest, indicating the influence of other factors in perception.
The overwhelming popularity of ratings in marketing and consumer research is well-documented, with Likert scales and semantic differential scales being among the most commonly used tools (Bruner, 2009). Many experiments and surveys heavily depend on these scales to gauge subjective consumer perceptions. For instance, in the advertising literature, it is common to ask respondents to rate the extent to which they perceive a brand, product, or advertisement as humorous (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), or involving (e.g., Stokburger et al., 2012). These perceptions are often crucial for understanding consumer behavior and preferences, shaping brand strategies and advertising campaigns. However, given the inherent subjectivity of such questions and the potential for personal bias, it is important to question how well ratings truly capture respondents' perceptions. Respondents may interpret the scale points differently or be influenced by contextual or emotional factors, leading to responses that are not entirely

reflective of objective reality. This subjectivity can distort the results, making it difficult to compare responses across individuals or groups accurately.

Consequently, there is a growing interest in understanding how ratings compare with alternative scaling approaches, such as rankings or pairwise comparisons, especially in terms of their ability to approximate "ground truth" or objective reality. Unlike ratings, which ask respondents to quantify their perceptions on a predefined scale, rankings require them to make direct comparisons between options. This distinction may reduce some of the ambiguity and bias associated with subjective ratings, but rankings also come with their own limitations, such as ipsative constraints that force comparisons between potentially incomparable items. Therefore, exploring how ratings perform in comparison to these alternative methods is essential for determining which approach provides the most accurate and reliable insights into consumer perceptions. By evaluating the strengths and weaknesses of each scaling method, researchers can better understand how well these tools reflect the real-world phenomena they aim to measure. This comparison is especially critical in fields like marketing and consumer research, where understanding subjective perceptions and attitudes can significantly impact business decisions, product development, and advertising effectiveness.

5. CONCLUSIONS

The present study contributes to the existing body of literature in several important ways. First, to the best of the author's knowledge, it represents one of the first empirical efforts to experimentally compare the usability of pairwise rankings and ratings in a controlled setting. By introducing physical objects into the analysis, it was possible to assess the gap between results provided by human perceptions (via different scales) and objective reality. The findings suggest that both ranks and ratings can offer a similar level of consistency in predicting real-world phenomena. Although each scale may function slightly differently depending on the type of object being evaluated, the overall performance of ranks and ratings appears to be comparable. Second, this study makes a significant contribution by integrating ranks into a field traditionally dominated by rating-based research. The results indicate that ranks and ratings are both informative and effective in predicting the correct order of various phenomena. Given the possibility that ranks produce less cognitive overload (as proposed by Yang and Chen, 2011) and fewer reporting biases (Yannakakis and Hallam, 2011), rankings may offer a more efficient and user-friendly alternative to ratings, especially in studies that demand high levels of attention and memory capacity from respondents. This suggests that ranks could be a valuable alternative in contexts where ratings might place undue cognitive demands on participants. Third, the study provides insights into the applicability of preference learning methods, specifically the *Preference Learning Toolbox* (PLT), in social research. Although artificial neural networks are extensively used in fields such as sales forecasting, investment estimations, and the gaming industry, they remain relatively uncommon in social sciences, particularly in areas such as advertising, branding, and marketing communications. The current study demonstrates that using PLT to predict rankings can be as effective and informative as traditional rating-based analyses, while also helping to avoid potential statistical issues associated with ratings (as highlighted by Ovadia, 2004). This opens new avenues for incorporating machine learning techniques into social research, offering the potential for more sophisticated and accurate modeling of human preferences. Despite these contributions, the experiment has certain limitations. It focuses primarily on how individuals perceive and report the physical characteristics of objects, comparing their subjective cognitions to objective reality. However, the study does not address affective judgments or subjective preferences. As such, future research is needed to test the usability of rankings and ratings in measuring not only cognitive evaluations but also emotional responses and less tangible phenomena, such as consumer satisfaction, liking, or anxiety. Additionally, the study's scope was limited to physical objects, meaning that future research should explore how these scaling methods perform in contexts involving more complex, abstract constructs. Future studies should also aim to explore a broader range of circumstances in which either rankings or ratings may be more effective. Understanding the specific contexts—such as different emotional states, cultural influences, or varying cognitive demands—where one scale type outperforms the other will be crucial for refining the use of these methods in social research. Furthermore, extending this work to more intricate phenomena could help solidify the practical benefits of rankings in empirical investigations, especially when dealing with subjective assessments that are difficult to quantify with traditional ratings.

REFERENCES

Brown, M.R., Bhadury, R.K., & Pope, N. (2010). The impact of comedic violence on viral advertising effectiveness. *Journal of Advertising, 39*(1), 49-66.

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

Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research, 69*(7), 2574-2584.

Cline, T.W., Altsech, M.B., & Kellaris, J.J. (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., Galekh, M., & Vonkeman, C. (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., & Simonson, I. (2003). The effect of forced choice on choice. *Journal of Marketing Research, 40*(2), 146-160.

Eisend, M., Plagemann, J., & Sollwedel, J. (2014). Gender roles and humor in advertising: The occurrence of stereotyping in humorous and non-humorous advertising and its consequences for advertising effectiveness. *Journal of Advertising, 43*(3), 256-273.

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

Geuens, M., & De Pelsmacker, P. (2017). Planning and conducting experimental advertising research and questionnaire design. *Journal of Advertising, 46*(1), 83-100.

Gosling, S.D., Augustine, A., Vazire, S., Holtzman, N., & Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking, 14*(1), 483-488.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*(12), 1217-1218.

Kim, E., Ratneshwar, S., & Thorson, E. (2017). Why narrative ads work: An integrated process explanation. *Journal of Advertising, 46*(2), 283-296.

Kim, Y., & Yoon, H. (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., & Gronlund, N. (2000). *Measurement and Assessment in Teaching*. Prentice-Hall.

Martinez, H.P., Yannakakis, G.N., & Hallam, J. (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., Dens, N., De Pelsmacker, P., & Goos, P. (2015). Consumer responses to different degrees of advertising adaptation: The moderating role of national openness to foreign markets. *International Journal of Advertising*,

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

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

Schwarz, N., Knäuper, B., Hippler, H., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570-582.

Stokburger-Sauer, N., Ratneshwar, S., & Sen, S. (2012). Drivers of consumer-brand identification. *International Journal of Research in Marketing, 29*(4), 406-418.

Voss, K.E., Spangenberg, E.R., & Grohmann, B. (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*(1), 870-877.

Wänke, M., & Schwarz, N. (1992). Comparative judgments: How the direction of comparison determines the answer. *ZUMA Conference Paper*.

Warren, C., & McGraw, P. (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., Geuens, M., & Baumgartner, H. (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*(1), 50-59.

Yang, Y.H., & Chen, H.H. (2011). Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(1), 762-774.

Yannakakis, G.N., & Hallam, J. (2011). Rating vs. preference: A comparative study of self-reporting. *Affective Computing and Intelligent Interaction, 6974*, 437-446.