Received: 7 February 2020 | Revised: 1 June 2020 | Accepted: 4 June 2020

DOI: 10.1111/joss.12593

ORIGINAL ARTICLE

Journal of Sensory Studies | WILEY

Estimates of relative acceptability from paired preference tests

Curtis R. Luckett | Sara L. Burns | Lindsay Jenkinson

Department of Food Science, University of Tennessee, Knoxville, Tennessee

Correspondence
Curtis R. Luckett, Department of Food Science, University of Tennessee, 2510 River Drive, Knoxville, TN 37996.
Email: cluckett@utk.edu

Abstract

The relative acceptability of a food or beverage is an important data point in many different scenarios. In humans, a nine-point category scale is typically used when the hedonic characteristics of several products are of interest. However, these scales are not consistently used by participants and therefore reliability is low. This article outlines the effectiveness of four techniques (mElo, Thurstonian modeling, Bradley–Terry, and Friedman) to calculate numerical values for products based upon their performance in a paired preference paradigm. In this study, acceptance data from four separate studies were compared to numeric scores constructed from a paired preference paradigm. In general, numeric ratings constructed from paired comparisons correlated very well with the mean overall liking ratings. The relationship between these acceptance ratings and numerical preference ratings did show to be somewhat dependent on the type of food product. This study serves as a guide to those looking to further quantify paired preference data.

Practical Applications: Paired preference tests are simple and easy to understand for those who struggle with test instructions and numerical scales. However, in a multisample paradigm, sensory scientists have overwhelmingly chosen to use rating scales for a variety of reasons. This study shows the effectiveness of alternative data analysis techniques to extract numerical values for relative acceptance based on paired preference tests. This information could be used in the implementation of paired preference tests to replace traditional affective testing in groups that do not perform well with scales (e.g., children) or as an alternate analysis method for paired preference tests.

1 | INTRODUCTION

It is clear there is a conceptual distinction between preference and acceptability, mainly preference tests do not indicate whether products are liked or disliked (Meilgaard, Carr, & Civille, 2016). Statements that differentiate liking and preference are often made on the basis of the relativity of preference, implicitly asserting that acceptability is a universal measure. However, acceptability scores are often interpreted in a relative manner based upon the product category. While the mean acceptability score of a product does give the researcher a measure of liking, without the context of the product category, it can be difficult to translate a liking rating to decisions regarding that product. The inclusion of products from across the product category in sequential monadic designs and historical results are commonly used to give acceptability scores context. We will refer to this concept as *relative acceptability*.

The nine-point hedonic scale has been the standard in measuring food and beverage hedonics (Peryam & Pilgrim, 1957). However, the difficulty of accurately and precisely rating perception using a scale has been widely documented (Coetzee & Taylor, 1996; Gay & Mead, 1992; O'Mahony, 1982; Simone & Pangborn, 1957; Wilkinson & Yuksel, 1997). First, bias is created by participants trying to self-monitor their response history and compare their ratings of previous samples to their perception of the current sample (Böckenholt, 2001). Second, consumers seem to struggle using the scale consistently and correctly. For example, relatively low

correlations (average of 0.552 across 11 foods) were found between acceptance scores of identical food products in the same session (Cardello & Maller, 1982). Additionally, it has been known for some time that the use of a nine-point hedonic scaling in food acceptance leads to some level of frustration by participants who like samples equally, but still prefer one (Simone & Pangborn, 1957). Beyond the typical consumer sensory participant in developed countries, certain populations also have issues using the scale effectively, such as young children and those without strong language skills (i.e., people using their second language or traditional cultures where literacy is minimal) (Coetzee & Taylor, 1996).

From a fundamental data perspective, there is also the issue of treating the data as unidirectional, when the scale itself is bipolar and the data commonly violates the assumption of normality (O'Mahony, 1982; Vie, Gulli, & O'Mahony, 1991). Recent work has shown the treatment of data from categorical scales, such as the nine-point hedonic scale, as continuous data systematically lead to a multitude of statistical errors (Liddell & Kruschke, 2018). Because of these issues, alternative methods to assess relative acceptability should be investigated.

We have chosen to assess four methods to give estimates of relative food/beverage acceptability from a paired preference paradigm. Paired comparison tasks are more sensitive, less reliant on cognitive or language capabilities, more informative (a single decision gives information about two samples), and show little bias across populations (Coetzee & Taylor, 1996; Hasegawa, Ishii, Kyutoku, Dan, & Rousseau, 2019; Laue, Ishler, & Bullman, 1954). Additionally, participant responses to paired preference tests have been reported to be stable over time (Halim, Sinaga, Hu, Sebastian, & O'Mahony, 2019).

As mentioned earlier, context is easily provided in direct scaling (e.g., historical scores, scores from competitive products). However, methods to give the context needed to estimate relative acceptability in a paired preference paradigm have not been well established. To gain a broad context in a paired preference paradigm, several representative pairs of samples would need to be assessed with a single product category. Unfortunately, paired preference paradigms are not commonly used for product sets containing several samples due to the dependence on each sample pair being seen together. The number of comparisons needed to ensure every pair of samples is compared against each other increases geometrically with the number of samples. For example, comparing six samples would necessitate 15 paired comparisons to ensure each participant compared each possible pair.

Historically, paired preference data are analyzed using simple comparisons using a binomial distribution (Amerine, Pangborn, & Roessler, 1965). More specifically, the data being analyzed are based on response methods, requiring each possible pair of samples to be compared to each other. The results of these analyses are the identification of a statistically significant preference between the two samples in question, not a measure of relative acceptability.

From within sensory and consumer science, there are existing methods to construct relative acceptability from a group of paired preference tests that have become popular. Both the r-index and Thurstonian modeling have been used to create estimates of performance in paired preference testing (Alfaro-Rodriguez, O'Mahony, & Angulo, 2005; Pipatsattayanuwong, Lee, Lau, & O'Mahony, 2001). The r-index approach has been well characterized in the literature (Lee & Van Hout, 2009), however, was not included in this study because of the unique requirement of asking participants the level of certainty in their judgment. Thurstonian models utilize signal detection theory to predict the degree of difference or preference between given samples, a value otherwise known as δ (Braun, Rogeaux, Schneid, O'Mahony, & Rousseau, 2004). This analytical method assumes that correlation exists among the observations, and has been adapted for sensorial discrimination and preference tasks to fulfill normal distribution assumptions (Cattelan, 2012). In most published applications, the Thurstonian approach calculates d' values between two samples. Meaning, if a series of paired preference tests were performed between three samples, the estimates (d') of preference between two samples are not affected by either of the samples' performance against the third sample. When the degree of relative acceptance is the goal, there is an increasing benefit to being able to extract information about a sample by using its performance in all judgments. From the work of Frederick Mosteller, we can expand the Thurstonian approach to give Thurstonian estimates of a sample that considers all of the paired comparisons in the sample set concurrently (Mosteller, 1951a; Mosteller, 1951b; Mosteller, 1951c). In this work, Mosteller details the construction of a value (S), which exists on a continuum, and is constructed using pooled performance against all other samples.

Similar to the Thurstonian method, the Bradley–Terry method postulates the existence of a preference continuum (Alfaro-Rodriguez et al., 2005; Gridgeman, 1955). Interestingly, while the Bradley–Terry method was developed to assess food preferences, it has been more widely adopted in the analysis and prediction of sporting events. The Bradley–Terry method models the association between judgments of paired samples in a sequence and is unique in the sense that it calculates values based upon a reference sample, considering all other scored samples in addition to the reference (Bradley & Terry, 1952; Gridgeman, 1955). Furthermore, there are Friedman-type analyses centered around rank order, which have also shown promise in analyzing a series of paired comparison tests (Carr, 1985; Friedman, 1937).

There are additional alternative methods of paired comparison data analysis that have been less frequently used in sensory applications. This article will explore the mElo method, which is derived from a popular chess player rating system (Elo, 1978). In the Elo method, each player is assigned a rating, and the fact that not all chess players would compete against each other is taken into account. When two players meet, the winner's rating is adjusted upward and the loser's rating is adjusted downward. The magnitude of the adjustment is directly related to how surprising the result is based upon the prior player ratings. While being theoretically sound and relatively simple, Elo ratings are not without fault. One of the most relevant criticisms centers around the fact that the sequence of the matches has a major influence on the ratings. To counter this problem, Neumann developed the mElo rating, in which the sequence of matches is randomized numerous times and the average of those values is calculated (2015). The number of randomizations is not standardized, but 100 to

1,000 randomizations have been used, with little to no advantages or disadvantages shown for either value (Clark, Howard, Woods, Penton-Voak, & Neumann, 2018).

While Bradley–Terry models have been used to quantify food preference in other species (Molloy & Hart, 2002), the use of Elo ratings on food preference data has not yet been applied. One of the few assumptions Elo and Bradley–Terry models make is that there is a "reasonable" degree of shared perception among the participants (Clark et al., 2018; Reid, Nixon, & Stevenage, 2014). It is well known that consumer liking for food products is not always in high agreement, justifying the need to test these methods in a food/beverage consumer setting.

The ability to accurately quantify relative acceptability from the paired preference test paradigm would be a key step forward in sensory testing for those who struggle with scale usage (e.g., children). Additionally, producing reliable numeric values from paired preference data can lead to the ability to implement other statistical tools commonly used, such as principal component analysis or clustering (Linander, Christensen, Cleaver, & Brockhoff, 2019). This study was designed to assess various methods of analyzing paired preference data and serves as a model for researchers looking for alternatives to nine-point hedonic scales in assessing relative acceptability.

2 | MATERIALS AND METHODS

2.1 | Stimuli

Four separate food and beverage products were used in this study. Each was chosen based upon its complexity and valence. Additionally, the degree of unbalance was modified across the four separate sensory tests. In some cases, participants judged all possible pairs of samples (e.g., coffee), while in others they saw just over half of the samples (e.g., cola).

2.2 | Beverages

Six commercially available cola-flavored sodas were purchased. Forty-five milliliters of each cola was served at 7°C in a 60 ml plastic cup. Additionally, three food-service cold brew coffees were provided by Pilot Flying J LLC. (Knoxville, TN). These coffees were prepared according to their instructions and served at 7°C in 60 ml plastic cups. Like the soda, 45 ml of each product was provided for evaluation. The participants were allowed to alter the coffee using sweetener and creamer to mimic how they usually drink their cold brew.

2.3 | Ice cream

Five commercially available vanilla ice creams were also purchased. Panelists were provided with 22 ml of sample in 100 ml lidded foam containers and 7.6 cm white plastic spoons.

2.4 | Pizza

Five commercially available frozen four-cheese pizzas were purchased and prepared according to package instructions. Each panelist received slices equaling one-twelfth of a pizza on 15 cm white Styrofoam plates.

2.5 | Participants

Participants were recruited using the University of Tennessee at Knoxville sensory consumer database. All participants in the database (n = 650) received an email announcing the testing date with a link to a screener. The screening survey was designed to select participants who were familiar with the product of interest. One hundred participants were randomly selected from those who signified a product consumption pattern. All 100 participants completed the coffee session, while 98 participants completed the soda sessions, 86 completed the vanilla ice cream series, and 91 completed both pizza sessions. All participants signed an informed consent form and were compensated for their time. This experiment was conducted according to the Declaration of Helsinki for studies on human subjects and approved by the University of Tennessee IRB review for research involving human subjects (IRB # 18-04342-XP).

2.6 | Procedure

For the cola-flavored soda, participants were asked to attend three experimental sessions. In the first session, the participants received the samples in a sequential monadic fashion, using randomized order, and were asked to rate their overall liking. In the subsequent two sessions, the participants saw the samples in a paired preference paradigm in which two samples were served together and the participant was asked to identify the samples they preferred (no preference options were not allowed). For the soda samples, the paired preference samples were served using a randomized, incomplete block design, in which participants randomly received 10 of the possible 15 pairs of samples. The session was accounted for in the experimental design; therefore, no participants received the same pair of samples twice.

For the cold brew coffee, participants completed the test in one session in which they performed three paired preference tests (all combinations) following rating the samples in a sequential monadic manner. The arrangement of the samples (left–right) was randomized across the participants. Additionally, a pool of 15, four-digit blinding codes were used for each sample, minimizing participant learning.

The same protocol was followed for both ice cream and pizza experiments. Participants attended two experimental sessions, completing the paired preference portion of the study during the first session using a randomized, incomplete block design receiving five of the possible 10 pairs in the same fashion as the cola samples. Participants rated samples in a random sequential monadic order during the

second session. Demographic information was collected in the final session for all. Lastly, all data were collected using Redjade (Redwood City, CA).

2.7 | Data analysis

All analyses were performed in R version 3.5.1. *S* values were calculated using the methods outlined by Mosteller (1951a), Mosteller (1951b), and Mosteller (1951c). More specifically, the proportion of times a sample was preferred was converted to its normal deviate using the *qnorm* function in R and averaged across all the comparisons (Mosteller, 1951c). mElo ratings were calculated using the EloChoice package developed by Neumann (2015). mElo ratings were calculated using 1,000 randomizations. Bradley–Terry values were calculated using the *BTm* function in the BradleyTerry2 R package (Turner & Firth, 2012). To assess the agreement of the four paired preference data analysis methods (Thurstonian modeling, mElo, Friedman, and Bradley–Terry) with nine-point hedonic ratings, Pearson correlations were run using the *psych* package in R (Revelle, 2019). Additionally, Kendall's rank correlation analysis was performed on the product rankings. For ease of interpretation, product rank sums were subtracted from 1,000. Lastly, since the numeric values constructed from paired preference tests are unitless and arbitrary, the values were standardized for further investigation. The standardized scores were then compared for dispersion through the range (minimum value to maximum value) and the distance to the next closest sample within each food-method combination. All data are available at https://osf.io/sgj6n/ (Luckett, 2018).

3 | RESULTS

The results for each experiment are presented in Table 1. For the Bradley–Terry method, the lowest-performing paired preference sample was used as the reference and received a value of 0.000. The nine-point hedonic ratings were significantly different ($p < .05$) across the samples in all four product groups. In general, the four methods used to analyze the responses to the paired preferences test performed similarly. When considering the rank order of the products, Kendall's tau revealed significant correlations between the product rankings for all data analysis methods ($p < .05$), or in other words, no significant differences in the rankings of the products were found. All four data analysis methods agreed on the lowest-performing samples for the four experiments. Further comparison of the product ranks can be seen in Figure 1. While the sample ranks were similar, we did find slight differences in the data across the four different food products.

TABLE 1 Standardized sample ratings

Sample	Overall liking	mElo	Bradley–Terry	Thurstonian	Friedman-style rank sum
Cola-flavored soft drinks					
S1	−0.33	−0.24	−0.67	−0.62	−0.74
S2	−0.50	−0.19	0.10	0.12	0.34
S3	0.71	0.45	0.47	0.44	0.65
S4	−1.23	−1.54	−1.39	−1.03	−1.47
S5	−0.25	−0.02	−0.06	−0.63	−0.09
S6	1.60	1.53	1.54	1.72	1.32
Cold brew coffee					
C1	0.85	0.64	0.53	0.53	0.53
C2	0.25	0.51	0.62	0.63	0.62
C3	−1.10	−1.15	−1.15	−1.15	−1.15
Ice cream					
I1	0.92	1.56	1.57	1.54	1.48
I2	−1.69	−0.01	0.08	0.05	0.04
I3	0.01	−0.08	−0.15	−0.01	0.04
I4	0.44	−0.25	−0.30	−0.35	−0.24
I5	0.32	−1.22	−1.19	−1.23	−1.32
Pizza					
P1	0.19	0.90	0.92	0.93	0.69
P2	−0.06	−0.60	−0.67	−0.68	−0.75
P3	0.77	0.31	0.56	0.56	0.56
P4	0.77	0.83	0.61	0.61	0.87
P5	−1.67	−1.43	−1.42	−1.41	−1.37

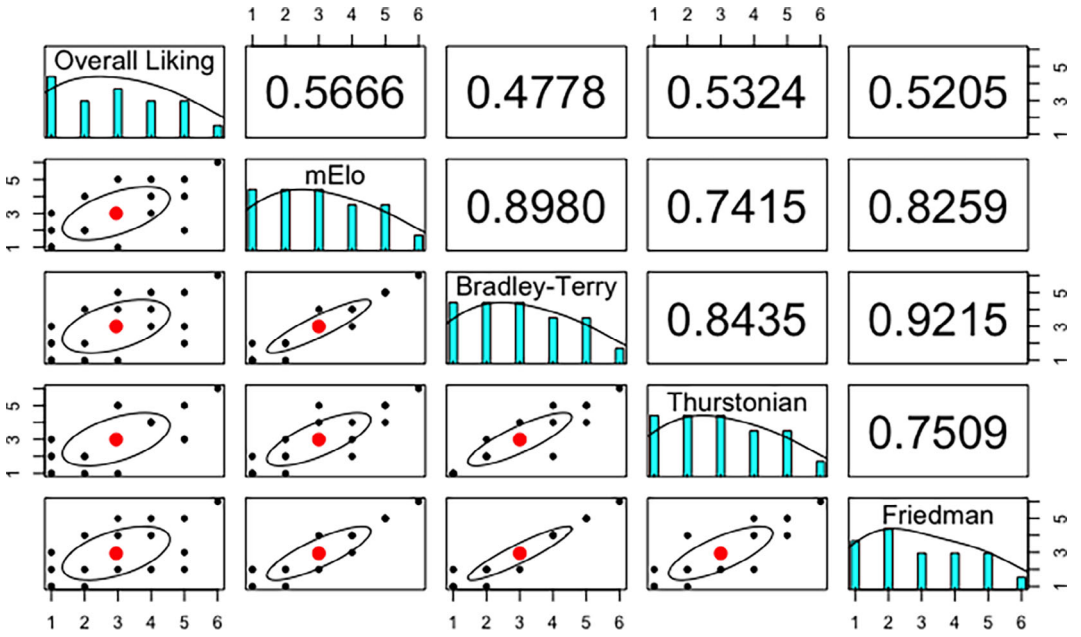

FIGURE 1 Correlation between mean overall liking rankings and relative acceptance rankings derived from paired preference testing. Numeric correlation coefficients are Kendall's τ. The red dot corresponds to the center of the data and the ellipse corresponds to 95% confidence interval. Axis values correspond to standardized relative acceptability ratings

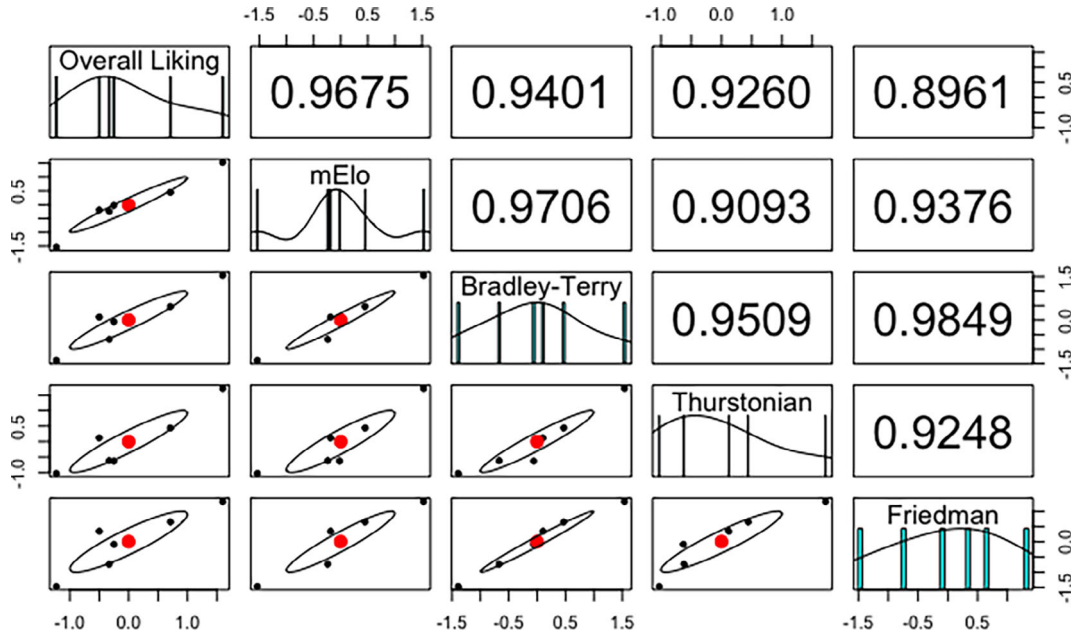

FIGURE 2 Correlation between mean overall liking scores and relative acceptance scores derived from paired preference testing for cola-flavored soft drinks. Numeric correlation coefficients are Pearson correlation coefficients. The red dot corresponds to the center of the data and the ellipse corresponds to 95% confidence interval. Axis values correspond to standardized relative acceptability ratings

In the cola-flavored soft drink samples, the mElo ratings were more highly correlated with the overall liking ratings, however, all methods of paired preference data analysis lined up very well with average nine-pt hedonic ratings (Figure 2). The highest correlation with overall liking scores was found with the mElo method (r = .9707), and the Friedman-style analysis was the least correlated to overall liking scores (r = .9255). Additionally, there was high agreement between the four analysis methods (r > .91).

The mean overall liking scores for pizza samples were also well correlated with the measures of relative acceptance produced from the paired preference data (Figure 3; r > .86). Of the paired preference analysis methods, no real differences were found in their relationship

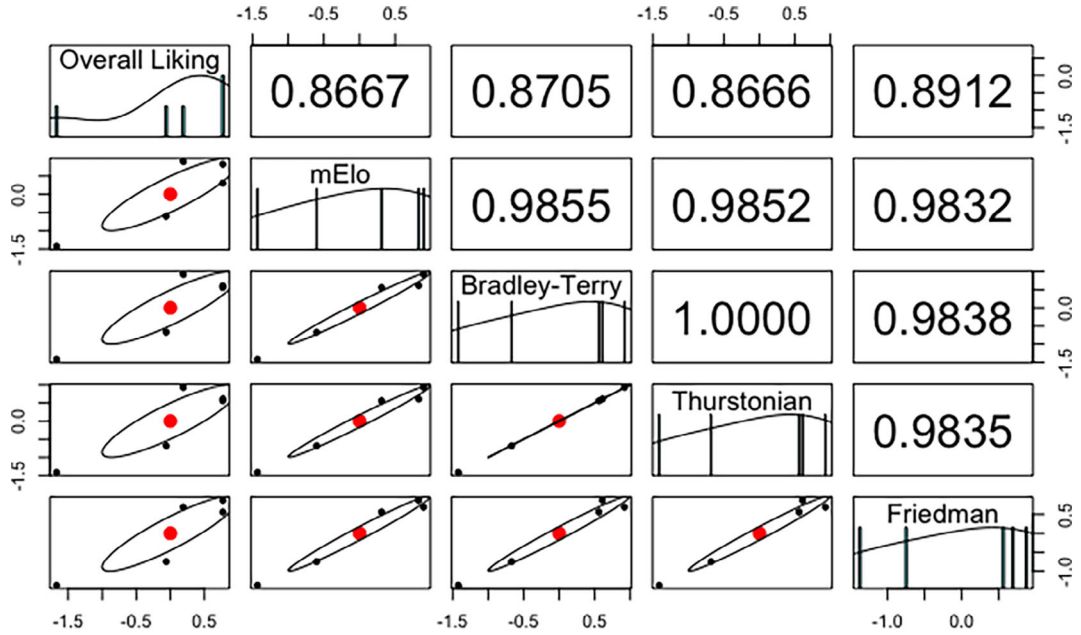

FIGURE 3 Correlation between mean overall liking scores and relative acceptance scores derived from paired preference testing for pizza samples. Numeric correlation coefficients are Pearson correlation coefficients. The red dot corresponds to the center of the data and the ellipse corresponds to 95% confidence interval. Axis values correspond to standardized relative acceptability ratings

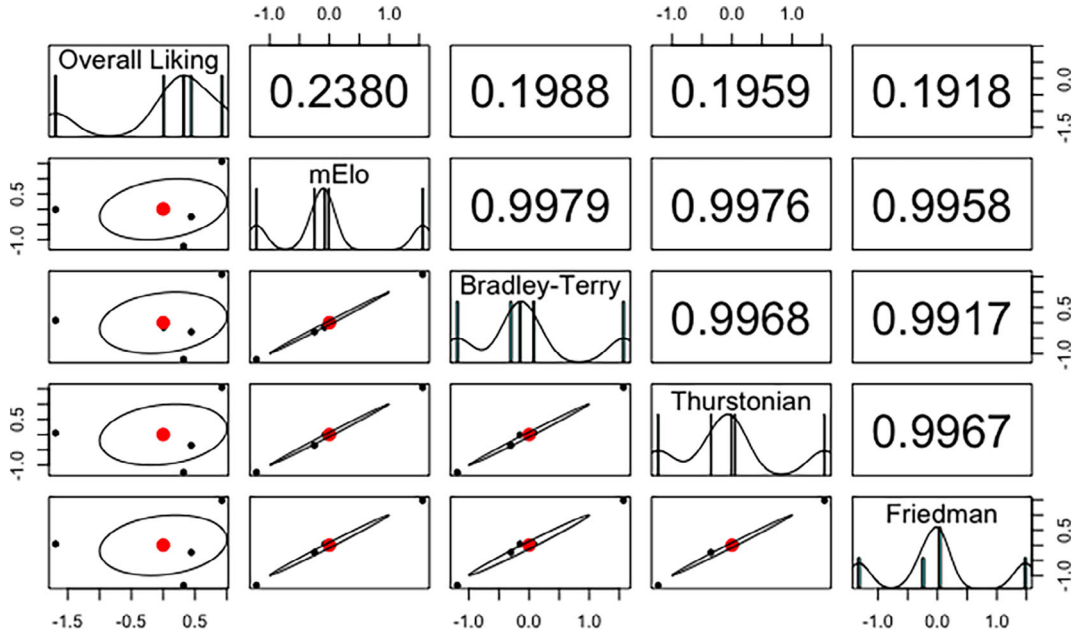

FIGURE 4 Correlation between mean overall liking scores and relative acceptance scores derived from paired preference testing for ice cream samples. Numeric correlation coefficients are Pearson correlation coefficients. The red dot corresponds to the center of the data and the ellipse corresponds to 95% confidence interval. Axis values correspond to standardized relative acceptability ratings

to mean overall liking scores. Additionally, a very high agreement was found among the paired preference methods ($r > .98$).

In regard to ice cream, we found that the paired preference results, no matter the analysis technique, did not reliably correlate to overall liking scores (Figure 4; $r < .28$). While no paired preference data analysis methods performed well, the values from the Thurstonian model had the highest correlation with mean overall liking scores ($r = .2728$).

The cold brew coffee samples displayed a strong relationship between mean overall liking scores and the measures of relative

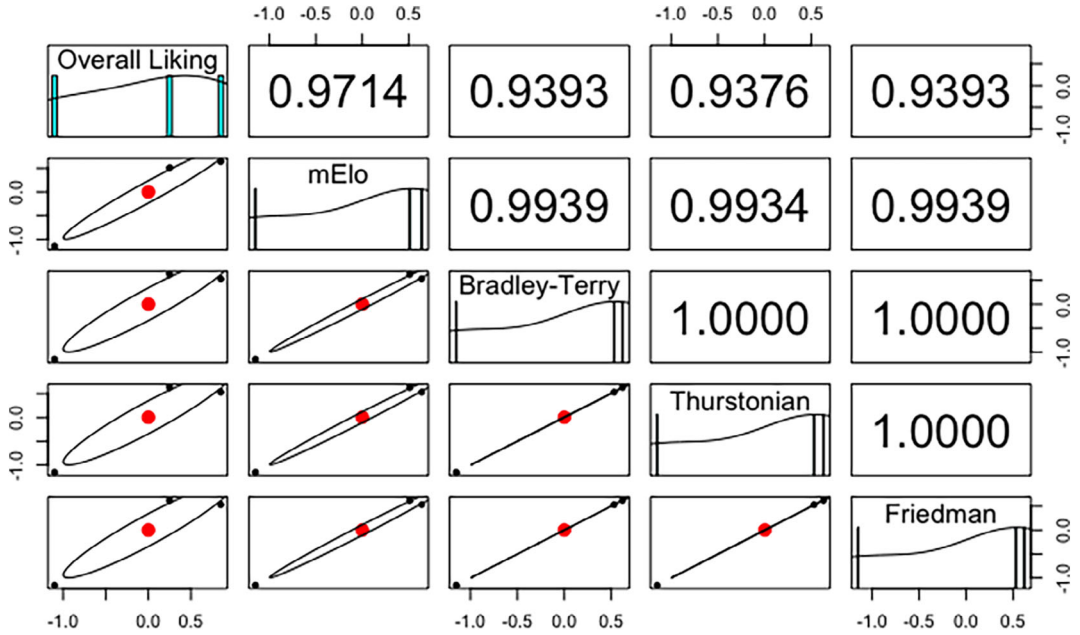

FIGURE 5 Correlation between mean overall liking scores and relative acceptance scores derived from paired preference testing for cold-brew coffee samples. Numeric correlation coefficients are Pearson correlation coefficients. The red dot corresponds to the center of the data and the ellipse corresponds to 95% confidence interval. Axis values correspond to standardized relative acceptability ratings

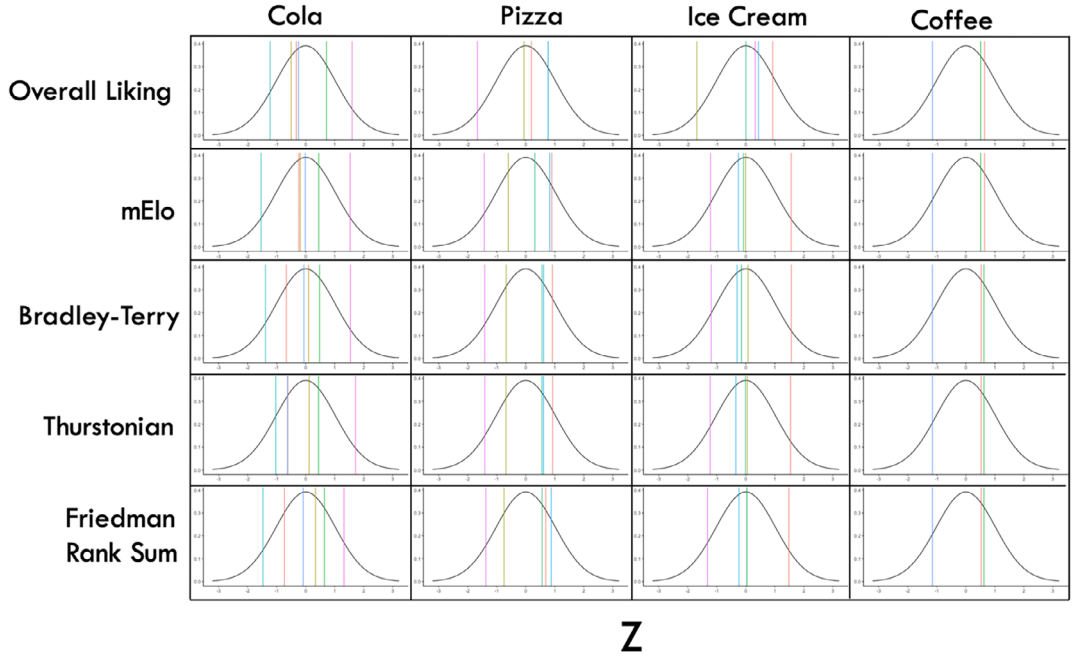

FIGURE 6 Standardized relative acceptance scores across four food products and five data analysis techniques. Each vertical line represents the mean relative acceptance score for a sample within that product set, analyzed with the method given in the row name

acceptance constructed from the paired preference paradigm (Figure 5; $r > .93$). The four separate measures of relative acceptance were very highly correlated to each other ($r > .99$). All the data for the coffee samples showed similar patterns of dispersion, with two of the three samples being within 1 *SD* of the mean.

To address the ability of the methods to differentiate between the samples, we observed minor differences (Figure 6). The range for all five methods was between 2.41 and 2.49 *SD*s from the mean score (Table 2). To further investigate possible differences in the dispersion of the samples, we calculated the absolute value of the distance from

TABLE 2 Range in standardized relative acceptability ratings (in *SD*s)

Overall liking		mElo	Bradley–Terry	Thurstonian	Friedman-style rank sum
Cola	2.83	3.07	2.93	2.75	2.80
Coffee	1.95	1.80	1.78	1.78	1.78
Ice cream	2.61	2.78	2.76	2.77	2.80
Pizza	2.44	2.33	2.34	2.34	2.25
Mean	2.46	2.49	2.45	2.41	2.41

TABLE 3 Distance to the nearest standardized rating of another sample (in *SD*s)

Overall liking		mElo	Bradley–Terry	Thurstonian	Friedman-style rank sum
Cola	0.47	0.52	0.52	0.39	0.52
Coffee	0.85	0.64	0.62	0.63	0.62
Ice cream	0.55	0.57	0.58	0.57	0.56
Pizza	0.42	0.46	0.38	0.38	0.34
Mean	0.57	0.55	0.53	0.49	0.51

each sample to the next closest sample (Table 3). Again, no notable differences were found across the analysis methods.

4 | DISCUSSION

This study demonstrates the use of four data analysis methods for obtaining estimates of the relative acceptability of samples from paired comparison data. When the results of the paired preference paradigm were similar to the results from the nine-point hedonic testing, all four methods of paired preference analysis provided measures of relative acceptance that were closely associated with the nine-point scale ratings. The mElo method created ratings most closely associated with hedonic scaling in two of the four foods, while Bradley–Terry and the Thurstonian method were the best performing for a single food. While the mElo method shows promise, it does have one unique aspect that could limit its usage. The mElo method does not have a measure of variance. To simplify the mathematics, Arpad Elo elected to negate that a player performs within a range of their true ability. This lack of variance is a known shortcoming of the Elo rating system and has since been expanded upon by others to include measures of variance (Glickman, 1999). Another known issue exists with the Thurstonian approach, which assumes a normal distribution and equal SD for each sample. Thurstone himself had noted that bimodal or skewed distribution could likely arise in a preference paradigm, invalidating his model (1957). However, in our studies, these issues did not seem to inhibit the performance of Thurstonian modeling.

It can also be stated that all four measures of relative acceptance constructed from paired preference tests are relatively similar. This is even more true when looking at the Thurstonian and Bradley–Terry models, which in some cases showed a perfect correlation (r = 1.0). From our results, it does not appear that the heterogeneity in food preference is ruinous to the use of models that are heavily dependent on the agreement, such as the Bradley–Terry and mElo methods. As stated earlier, one of the main assumptions made by both models is that there is a reasonable agreement between responses. As a caveat to the above statement, the lack of agreement between paired preference testing and nine-point hedonic ratings for the ice cream could be caused by preference segments. Future work should employ the analysis methods outlined by Bradley (1954) on a larger group of paired preference data sets to assess this point.

As mentioned earlier, when looking at the scores themselves, the construction of numeric relative acceptance scores from paired preferences seems to be more stratified. Higher levels of stratification within the scores could be used as evidence for higher test sensitivity, or a better ability to detect differences in relative acceptance. It would not be surprising to observe differences in test sensitivity due to reported hidden preferences, where participants like samples equally but still have a clear preference (Simone & Pangborn, 1957). However, when using standardized scores, statements about the superiority of any of the methods in detecting relative acceptance cannot be made. When looking across all five analysis methods and four foods, no clear superiority in terms of sensitivity is observed.

While this study was designed to expose the effectiveness of various methods to analyze paired preference data in assessing the relative acceptance of food, the relatively high correlations between relative acceptance scores calculated from paired preference tests and mean nine-point hedonic scaling provides some evidence that acceptance and preference are not notably different within a product category. Acceptance is roughly defined as the degree to which a product is liked, while preference refers to the preference for a particular sample within a set (Moskowitz, 2005). While acceptance through hedonic scaling and preference are notably different, this article shows that both methods can be used to estimate relative acceptance. Further research should be performed to better understand how consumers differentiate between preference, liking, and relative acceptance.

The logical next step for the methodologies presented in this article is sensory testing in children. Children are one of the primary

groups of food consumers that the traditional nine-point hedonic scale does not service well. Future studies should seek to replicate this work on children using scales targeted for children, such as the emoji scale (Swaney-Stueve, Jepsen, & Deubler, 2018). This study shows that mElo and Bradley–Terry ratings are highly correlated to nine-point liking scores. Therefore, if the desire is to measure the same concept as adults using the nine-point hedonic scale, calculating hedonic ratings from the paired preference of children appears to be an effective option. Paired preference tests have been shown to be readily understood by children as young as 4 years old, while children have not been shown to understand hedonic scaling until 7 years old (Guinard, 2000; Kimmel & Guinard, 1994). Further studies should directly compare the use of paired preference tests to smiley-face and emoji scales in children.

4.1 | Limitations

Just as there are published accounts of the shortcomings of the nine-point hedonic scale, paired preferences also have been shown to have inherent flaws (Marchisano et al., 2003). Most notably, paired preference tests are reported to have a hidden demand characteristic, which is a perceived demand of a task assumed by the participants. For paired preference tests, it is possible that participants would assume the samples are unique based upon the request to give a preference. This has led to a high frequency of preference responses when consumers receive identical samples (MacFie, 2007). However, *false preferences* have also been reported using nine-point hedonic scales (Villegas-Ruiz, Angulo, & O'Mahony, 2008). More specifically, only about 20% of participants gave putatively identical foods the same hedonic rating.

There are also more practical considerations that would need to be made when deciding whether to move forward with a paired preference paradigm. First, unlike in sequential monadic affective testing, paired preference testing does not offer the possibility of assessing the acceptance of individual attributes. Depending on the number of samples in the set, this could require panelists to assess a larger number of samples, particularly when evaluating all possible pairs, increasing the testing time and the chance of panelist fatigue. The increased number of sample pairs could also put a strain on test logistics if the sample quantity is limited.

5 | CONCLUSIONS

The use of numeric ratings produced from a series of paired preference tests to assess product relative acceptance in a multisample situation is a viable alternative to traditional nine-point hedonic scaling, when relative acceptability is of interest. Similarly, the percept of relative acceptability can be measured by either hedonic scaling or a series of paired preference tests. More practically, this study provides the framework for calculating valid numeric relative acceptability ratings from paired preference data.

ACKNOWLEDGMENTS

The authors would like to thank Christof Neumann for his guidance and construction of the EloChoice R package.

CONFLICT OF INTEREST

The authors declare that they do not have any conflict of interest.

ORCID

Curtis R. Luckett https://orcid.org/0000-0003-3955-7474

REFERENCES

Alfaro-Rodriguez, H., O'Mahony, M., & Angulo, O. (2005). Paired preference tests: d′ values from Mexican consumers with various response options. *Journal of Sensory Studies*, *20*(3), 275–281. https://doi.org/10.1111/j.1745-459X.2005.00018.x

Amerine, M. A., Pangborn, R. M., & Roessler, E. B. (1965). *Principles of sensory evaluation of food*. New York/London: Academic Press.

Böckenholt, U. (2001). Thresholds and Intransitivities in pairwise judgments: A multilevel analysis. *Journal of Educational and Behavioral Statistics*, *26*(3), 269–282. https://doi.org/10.3102/10769986026003269

Bradley, R. A. (1954). Incomplete block rank analysis: On the appropriateness of the model for a method of paired comparisons. *Biometrics*, *10*(3), 375–390.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Braun, V., Rogeaux, M., Schneid, N., O'Mahony, M., & Rousseau, B. (2004). Corroborating the 2-AFC and 2-AC Thurstonian models using both a model system and sparkling water. *Food Quality and Preference*, *15*(6), 501–507. https://doi.org/10.1016/j.foodqual.2003.10.002

Cardello, A. V., & Maller, O. (1982). Relationships between food preferences and food acceptance ratings. *Journal of Food Science*, *47*(5), 1553–1557. https://doi.org/10.1111/j.1365-2621.1982.tb04981.x

Carr, B. T. (1985). Statistical models for paired-comparison data. In *American society for quality control 39th congress transactions* (pp. 295–300). Baltimore, MD: American Society for Quality.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, *27*(3), 412–433.

Clark, A. P., Howard, K. L., Woods, A. T., Penton-Voak, I. S., & Neumann, C. (2018). Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PLoS One*, *13*(1), e0190393. https://doi.org/10.1371/journal.pone.0190393

Coetzee, H., & Taylor, J. R. N. (1996). The use and adaptation of the paired-comparison method in the sensory evaluation of hamburger-type patties by illiterate/semi-literate consumers. *Food Quality and Preference*, *7*(2), 81–85. https://doi.org/10.1016/0950-3293(95)00039-9

Elo, A. E. (1978). *The rating of chessplayers, past and present*, New York, NY: Arco Pub.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701.

Gay, C., & Mead, R. (1992). A statistical appraisal of the problem of sensory measurement. *Journal of Sensory Studies*, *7*(3), 205–228. https://doi.org/10.1111/j.1745-459X.1992.tb00533.x

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(3), 377–394.

Gridgeman, N. T. (1955). The Bradley–Terry probability model and preference testing. *Biometrics*, *11*(3), 335–343. https://doi.org/10.2307/3001772

Guinard, J.-X. (2000). Sensory and consumer testing with children. *Trends in Food Science & Technology*, *11*(8), 273–283. https://doi.org/10.1016/S0924-2244(01)00015-2

Halim, J., Sinaga, W. S., Hu, R., Sebastian, A., & O'Mahony, M. (2019). What do laboratory preference tests tell us about real life (operational) preferences: A preliminary investigation. *Food Quality and Preference*, *76*, 60–70.

Hasegawa, Y., Ishii, R., Kyutoku, Y., Dan, I., & Rousseau, B. (2019). Biases in paired preference tests: Cross-cultural comparison of Japanese and American consumers. *Journal of Sensory Studies*, *34*(3), e12498. https://doi.org/10.1111/joss.12498

Kimmel, S. A., & Guinard, J. X. (1994). Sensory testing with young children. *Food Technology*, *48*, 92–99.

Laue, E. A., Ishler, N. H., & Bullman, G. A. (1954). Reliability of taste testing and consumer testing methods. 1. Fatigue in taste testing. *Food Technology*, *8*(9), 387–388.

Lee, H. S., & Van Hout, D. (2009). Quantification of sensory and food quality: The R-index analysis. *Journal of Food Science*, *74*(6), R57–R64.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Linander, C. B., Christensen, R. H. B., Cleaver, G., & Brockhoff, P. B. (2019). Principal component analysis of d-prime values from sensory discrimination tests using binary paired comparisons. *Food Quality and Preference*, *81*, 103864.

Luckett, C. (2018). *Data*. Retrieved from https://doi.org/10.17605/OSF.IO/B2J7E.

MacFie, H. (Ed.). (2007). Consumer-led food product development. CRC Press. Boca Raton, FL.

Marchisano, C., Lim, J., Cho, H. S., Suh, D. S., Jeon, S. Y., Kim, K. O., & O'Mahony, M. (2003). Consumers report preferences when they should not: A cross-cultural study. *Journal of Sensory Studies*, *18*(6), 487–516. https://doi.org/10.1111/j.1745-459X.2003.tb00402.x

Meilgaard, M.C., Carr, B. T., & Civille, G. V. (2015). *Sensory evaluation techniques*, Boca Raton, FL: CRC Press.

Molloy, L., & Hart, J. A. (2002). Duiker food selection: Palatability trials using natural foods in the Ituri Forest, Democratic Republic of Congo. *Zoo Biology*, *21*(2), 149–159. https://doi.org/10.1002/zoo.10021

Moskowitz, H. R. (2005). Thoughts on subjective measurement, sensory metrics and usefulness of outcomes. *Journal of sensory studies*, 20(4), 347-362.

Mosteller, F. (1951a). Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, *16*(1), 3–9.

Mosteller, F. (1951b). Remarks on the method of paired comparisons: II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika*, *16*(2), 203–206.

Mosteller, F. (1951c). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. I. *Psychometrika*, *16*(2), 207–218.

Neumann, C. (2015). EloChoice: Preference rating for visual stimuli based on elo ratings (version 0.29).

O'Mahony, M. (1982). Some assumptions and difficulties with common statistics for sensory analysis. *Food Technology*, *36*(11), 75–82.

Peryam, D. R., & Pilgrim, F. J. (1957). Hedonic scale method of measuring food preferences. *Food Technology*, *11*, *Suppl*, 9–14.

Pipatsattayanuwong, S., Lee, H. S., Lau, S., & O'Mahony, M. (2001). Hedonic R-index measurement of temperature preferences for drinking black coffee. *Journal of Sensory Studies*, *16*(5), 517–536.

Reid, D. A., Nixon, M. S., & Stevenage, S. V. (2014). Soft biometrics; human identification using comparative descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(6), 1216–1228. https://doi.org/10.1109/TPAMI.2013.219

Revelle, W. (2019). Psych: Procedures for psychological, psychometric, and personality research (version 1.9.12.31).

Simone, M., & Pangborn, R. M. (1957). Consumer acceptance methodology-one vs 2 samples. *Food Technology*, *11*(9), A25–A29.

Swaney-Stueve, M., Jepsen, T., & Deubler, G. (2018). The emoji scale: A facial scale for the 21st century. *Food Quality and Preference*, *68*, 183–190.

Turner, H., & Firth, D. (2012). Bradley–Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, *48*(9).

Vie, A., Gulli, D., & O'Mahony, M. (1991). Alternative hedonic measures. *Journal of Food Science*, *56*(1), 1–5. https://doi.org/10.1111/j.1365-2621.1991.tb07960.x

Villegas-Ruiz, X., Angulo, O., & O'Mahony, M. (2008). Hidden and false "preferences" on the structured 9-point hedonic scale. *Journal of Sensory Studies*, *23*(6), 780–790. https://doi.org/10.1111/j.1745-459X.2008.00184.x

Wilkinson, C., & Yuksel, D. (1997). Modeling differences between panelists in use of measurement scales. *Journal of Sensory Studies*, *12*(1), 55–68. https://doi.org/10.1111/j.1745-459X.1997.tb00053.x

How to cite this article: Luckett CR, Burns SL, Jenkinson L. Estimates of relative acceptability from paired preference tests. *J Sens Stud*. 2020;35:e12593. https://doi.org/10.1111/joss.12593