Food Quality and Preference 127 (2025) 105442

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

# Measuring overall difference from a combination of attribute ratings with the many-facet Rasch model

Nnenna C. Ariakpomu *, Melvin J. Holmes, Peter Ho

*School of Food Science and Nutrition, University of Leeds, Woodhouse Lane, LS2 9JT Leeds, United Kingdom*

ARTICLE INFO

*Keywords:*
Many-Facet Rasch model
Difference-from-control
Attribute rating
Overall sensory difference
Friedman test

ABSTRACT

The Total Intensity Measure (TIM) approach offers an innovative solution for quality control by combining ratings of individual sensory characteristics into a single measure using the Many-Facet Rasch Model (MFRM). While the traditional Difference-From-Control (DFC) test is simple and useful for comparing products against a standard, it requires significantly more samples when examining a larger number of products unlike in attribute difference tests. This study aims to determine if the TIM method can serve as an alternative to the DFC method when comparing samples against a control. An untrained panel ($n = 67$) evaluated three UK commercial brands of Jaffa cakes using attribute difference and DFC tests. Assessors evaluated samples in triplicates according to each test's procedure on two different days. Friedman tests on the DFC scores compared to Rasch-produced measures of the combined attributes both showed significant differences between samples ($P < 0.01$). Pairwise comparisons with a control ($\alpha = 0.01$) for the DFC showed only one brand was different from the control, while the TIM showed that both brands were different from the control. Additionally, the Many-Facet Wright map showed the degree to which each attribute contributed to the overall difference. Of the five attributes evaluated, Sweetness and Orange flavor contributed the most followed by Cocoa flavor. Milky flavor and Saltiness did not contribute significantly, highlighting that while all attributes were assessed, only certain ones had a notable impact on the overall product differences. The proposed method is potentially beneficial to sensory analysts in obtaining better diagnostic information to support decisions about product differences.

## 1. Introduction

For quality control and quality assurance purposes (such as shelf life testing, product benchmarking, or accessing batch-to-batch differences), several researchers (Costell, 2002; Muñoz et al., 1992; Rogers, 2017) suggest the difference-from-control (DFC) test as one of the best methods to employ. It is simple to use and has the unique feature of assessing, not only the existence of an overall difference between product samples but also the magnitude of the differences relative to a product standard (a carefully selected control or reference sample). This quantitative assessment gives the DFC an advantage over other overall difference tests like the Triangle and Duo-Trio tests, which only provide binary data indicating whether a sample is different or not. Additionally, rather than requiring assessors to skim through multiple samples simultaneously to detect a difference (as in the Triangle test), evaluating samples in comparison to a control is less cognitively demanding, reducing the complexity of the test. However, the use of reference samples presents a challenge in itself, inducing expectation bias, as assessors may unconsciously anchor their ratings to perceived intensities in the reference sample (Lawless & Heymann, 2010; Rogers, 2017).

Although the DFC test provides a quantitative assessment to support product difference decisions, it shares a limitation with other overall difference tests in that it does not reveal the specific attributes causing the perceived differences. Several studies have addressed this by incorporating additional methods with the DFC. Rogers (2017) suggests including a comment section to gather insights into the possible cause of perceived differences. Compusense (2020), in their white paper on quality control with the DFC, demonstrated the use of follow-up check-all-that-apply (CATA) questionnaires to improve manufacturers' chances of identifying product faults, and Higgins and Hayes (2020) combined CATA questions and an open-ended comment box to further characterize differences in beer samples. However, the depth of insights obtained remains limited, as only qualitative data is generated regarding the presence or absence of an attribute, failing to capture nuances about which attributes were perceived more strongly as responsible for the product differences.

* Corresponding author.
*E-mail addresses:* fs17ncu@leeds.ac.uk (N.C. Ariakpomu), m.j.holmes1@leeds.ac.uk (M.J. Holmes), p.ho@leeds.ac.uk (P. Ho).

https://doi.org/10.1016/j.foodqual.2025.105442
Received 20 August 2024; Received in revised form 12 January 2025; Accepted 19 January 2025
Available online 22 January 2025

Attribute rating (AR) tests, on the other hand, require assessors to rate the perceived intensities of specific attributes, yielding quantitative data. It is a valuable aspect of descriptive sensory profiling tests used to quantify the identified sensory characteristics of a product using attribute intensity rating scales. However, descriptive sensory profiling can be time-consuming and expensive due to its complexity. Assessors must generate a sensory lexicon (a comprehensive list of sensory attributes relevant to the product category) and undergo continuous training to ensure and maintain their sensory acuity and consistency.

The need to accelerate new product development and deliver faster innovations that meet consumer expectations has resulted in the development of rapid sensory profiling methods. The Rate-all-that-apply (RATA) test, for instance, can use untrained assessors or consumer panels to rate the intensities of only the sensory attributes they perceive to be present in the samples based on a predefined list of sensory descriptors. It is essentially an AR test providing quantitative data and has been reported to improve sample discrimination compared to CATA (Ares et al., 2014; Reinbach et al., 2014).

The AR test offers another advantage over the DFC: samples are evaluated independently, rather than in comparison to a control. In contrast, comparing samples to a control means the DFC can be resource-intensive, requiring a larger number of samples when testing multiple products. Assessors may experience fatigue from tasting numerous samples during a single test session. Even when samples are evaluated across multiple sessions, it still demands additional time commitment. However, unlike the DFC, AR tests cannot directly quantify an overall difference between samples; instead, insights on product differences need to be captured through complex multivariate statistical analysis.

A Rasch approach can address some of the limitations of the DFC test by requiring fewer samples and statistical analyses without compromising diagnostic information. In this approach, consumer panels rate the intensities of predefined attributes. Next, a Many-Facet Rasch Model (MFRM) is fitted to combine these attribute ratings and estimate a latent variable representing the overall difference in intensity as has been done previously for overall liking assessments (Ho, 2019). The resulting Rasch measures are then used for univariate statistical analysis to determine overall differences between samples and a prescribed control. An additional benefit is that information related to which attributes contribute more to the measure of overall differences can be determined.

Rasch analysis is a statistical method used for analyzing categorical data in surveys or assessments. Its goal is to measure unobservable or latent variables (such as ability or overall attitudes) using a combined set of items (like questions in a survey). The probability of a correct response to an item is modeled as a logistic function of the relative distance between a person's location and an item's location on a common linear scale (Bond et al., 2021; Boone et al., 2014). In this context, *Persons* refer to the objects of measurement (such as respondents in the survey) while *items* correspond to the questions in the survey. For sensory evaluation, consider persons as the samples being evaluated and items as the questions in the sensory evaluation questionnaire.

The Many-Facet Rasch model (Linacre, 1994) extends the Rasch model allowing for simultaneous analysis of multiple facets that represent different sources of variability such as different samples, attributes, and assessors (raters). With this approach, responses from sample ratings can provide more diagnostic detail about each sample's performance, each assessor's behavior (particularly regarding the use of the rating scale), the intensity of each sensory attribute, and how much it contributes to the sample performance, and the functioning of the rating scale used for evaluating the samples; all within a single analysis. Several studies (Alvarez & Blanco, 2000; Faye et al., 2013; Ho, 2019; Thompson, 2003) have demonstrated the potential of Rasch modeling in the sensory evaluation of foods. A detailed description of the Rasch models and how they are interpreted in a sensory evaluation setting can be found in (Ho, 2019). Additionally, readers seeking a deeper understanding of the principles behind the family of Rasch models can refer to Bond et al. (2021); Boone et al. (2014); and Eckes (2023).

### 1.1. Aim of the Study

This study proposes a method that measures overall sample differences using a set of sensory attribute ratings which can be combined into a single Total Intensity Measure (TIM) with the Many-Facet Rasch Model (MFRM). The aim was to evaluate whether the TIM approach could be equally as effective as the DFC test when measuring overall differences between products and comparing differences against a control product. Where, the TIM method provides the added benefit of showing the relative contribution of each attribute to the overall difference and so would provide analysts with better diagnostic information regarding attributes that are driving product differences.

## 2. Materials and methods

### 2.1. Samples

Jaffa cakes were used for the study. These are sponge cakes with three layers: a sponge base, an orange-flavored jam center, and a chocolate topping. This sample choice was based on selecting a product with similar taste attributes (orange flavor, chocolate flavor, sweetness, saltiness, and milky flavor) as samples used in a previous study (Gill et al., 2024) and how findings relate to a larger study on measuring sensory differences using Rasch modeling. The samples were also required to be similar in visual appearance and texture but taste different, as taste attributes were the focus of the study.

Three samples, comprising one premium and two store-brand Jaffa cakes, were selected based on informal tasting sessions conducted by the research team, as well as back-of-pack label information. These samples were purchased from major supermarkets in the United Kingdom and differed in their nutritional and ingredient composition, including cocoa, orange flavoring, milk, salt, and sugar contents, as assessed from the label information. While the store brands were very similar in appearance, the premium brand had a slightly different shape (Fig. B1). One of the store-brand samples was selected as the reference for the DFC test. The samples were stored in odor-free, airtight, plastic containers at room temperature (20 ± 3 °C) until they were ready to be presented.

### 2.2. Participants

Ethical approval was granted by the University of Leeds Faculty of Environment ethics committee before commencing the sensory study.

Participants ($n$ = 67) included residents of Leeds, mostly staff and students from the University of Leeds who were recruited through emails, advertisement posters and word of mouth referrals. They were screened based on the following criteria: between 18 and 65 years old, with no chronic health conditions, no allergies or food intolerances to the ingredients in the Jaffa cake samples, were not on any routine medication (except contraceptives) nor on a restricted diet, were neither pregnant nor lactating, and their availability to attend two 1-hour-long sensory test sessions, within one month and with a minimum of four days between sessions.

While a trained panel is typically ideal for quality control testing scenarios, an untrained panel was used in this study to explore how the TIM approach performs in varying levels of sensory expertise. In a previous yet-to-be published study where the performance of trained ($n$ = 7) and untrained assessors ($n$ = 24) was compared, Rasch quality control metrics revealed some assessors in the untrained panel with qualities (e.g., consistency across replicates) indicating potential for trained-level performance. The intention was to investigate whether the performance patterns observed in the prior study could be replicated with a broader group of untrained assessors. The final untrained panel comprised 43 females (64 %) and 24 males (36 %) aged between 18 and 54 years old. All participants gave informed consent and were

incentivized for their participation.

### 2.3. Sensory evaluation

A randomized complete block design (RCBD) and William's Latin Square, as described by Næs et al. (2010), were used for the sensory experiments to account for order effects and other possible sources of variation. Each assessor participated in two separate sessions, one for the DFC test and another for the AR test, with a minimum interval of 4 days between each session. To minimize expectation biases (Lawless & Heymann, 2010), half of the participants completed the AR test session first, while the other half started with the DFC test. Attendance was balanced for the time of day and for which of the two tests they first completed. In each session, three samples were presented: for the AR test, samples were presented monadically (one at a time), while for the DFC, the samples were presented in pairs consisting of a test sample and the reference sample. Each sample was evaluated three times, making a total of 9 evaluations for AR and 18 for the DFC. All samples were served at room temperature (20 ± 3 °C) on 15 cm white paper plates labeled with random 3-digit codes. The reference sample for the DFC was labeled "R." Sensory evaluation was conducted in individual booths under white light at the sensory laboratory of the School of Food Science and Nutrition, University of Leeds. Data were collected using RedJade sensory software. (Redjade Software Solutions, 2023).

#### 2.3.1. Testing procedures

The Difference-from-control (DFC) test followed the procedure described by Meilgaard et al. (2016). Assessors were informed that some coded test samples might be the same as the reference and were instructed to taste each sample by taking a semi-circle shaped (half) bite. This instruction was necessary because Jaffa cakes are designed with the layer of orange-flavored jam centrally positioned on one side of the sponge base, which is then covered with a layer of chocolate. Without this guidance, assessors might only take a bite from the edge, missing the orange-flavored center and compromising the uniformity of the sample evaluation. They were instructed to first taste the sample labeled "R", then taste the coded test sample, assess the overall difference between them and rate the size of difference perceived. Assessors used a labeled 7-point categorical difference scale (0–6), where 0 = no difference, 1 = barely detectable difference, 2 = slight difference, 3 = moderate difference, 4 = large difference, 5 = very large difference, and 6 = extremely different, to rate the size of differences between a coded test sample and the reference sample (R).

For the Attribute Rating (AR) test, assessors rated the perceived intensities of five taste attributes: orange flavor, sweetness, cocoa flavor, milky flavor, and saltiness. As mentioned in section ***2.1. Samples***, these attributes were selected based on a preliminary study involving a product with a similar taste profile, where a trained panel from a global chocolate manufacturing company identified these attributes for orange-flavored chocolate spreads. The same attributes were used in this study to explore the method with a different product. Assessors were asked to taste each sample and rate how strong each of the five attributes were. All the attributes were presented on the same page of the questionnaire, but the order was randomized for each sample and assessor, as suggested by Ares et al. (2014) attempting to reduce errors of habituation, logic and halo effect (Lawless & Heymann, 2010). An 8-point categorical intensity scale ranging from 0 to 7 with labels adapted from the Labeled Magnitude Scale (Green et al., 1996) was used. The intensity labels were 0 = none, 1 = barely detectable, 2 = weak, 3 = moderate, 4 = strong, 5 = very strong, 6 = extremely strong, and 7 = strongest imaginable oral sensation. The inclusion of the "none" label represented the 0 point on the LMS, while adding "extremely strong" seemed an appropriate intensity rating between "very strong" and "strongest imaginable sensation" for use in a labeled categorical scale where there is no continuous line to mark intensity estimates, unlike the LMS. Additionally, the term "extremely" has been used in other category-ratio intensity scales, such as the Borg scale and its modifications (Borg, 1982; Borg & Kaijser, 2006).

Assessors were provided with a cup of water to cleanse their palate between sample evaluations and given breaks between replicates (5 min for the DFC and 10 min for the AR test) to minimize sensory fatigue and memory bias, respectively.

### 2.4. Defining the construct of overall difference as a composite of attribute intensity ratings

The theoretical development of measurement instruments for Rasch analysis must be carefully developed to capture the parameters of the latent variable to be measured (Boone, 2016). For this study, the construct modeling framework as described by (Ho, 2019) was adapted for defining overall difference as a latent variable estimated from a combination of attribute intensity ratings. (See Fig. 1.)

***Step 1:*** How would the theoretical construct of overall difference be defined? Sensory attributes that represent the sensory characteristics and modalities of the choice sample should be identified and used to capture different amounts of the latent variable of Overall difference. A minimum of 3–5 sensory attributes are recommended to ensure sufficient variability in the data and allow for the Rasch model to effectively separate the effects of the different facets.

In this study, the DFC ratings assessed overall product differences. To ensure comparability with the AR test, attempts were made to ensure that all sensory characteristics, except for taste, were consistent across the samples. This was done to minimize the influence of other sensory modalities on the perception of overall difference. The intensities of the selected taste attributes were hypothesized to represent components of the overall difference construct for the Jaffa cake samples.

**Step 2:** Five taste attributes were selected (orange flavor, sweetness, cocoa flavor, milky flavor, and saltiness) and survey questions were developed based on those attributes for assessors to rate their perceived intensities for each sample. For example, 'How strong is the orange flavor for sample xxx?' These then constituted the item measures for the model.

***Step 3:*** The 8-point category rating scale (described in ***2.3.1 Testing procedures*** for the AR test) representing the levels of possible perceived intensities was used by the panel of assessors - the raters for the Rasch model.

***Step 4:*** Observations were collected as attribute intensity ratings for each sample.

***Step 5:*** A MFRM with four facets comprising, assessors (the raters), samples (the persons), attributes (the items), and repetition (see **…Eq.** 1), was fitted. The resulting TIM were then used for ANOVA and multiple comparison tests.

***Step 6:*** The Rasch model's Wright map visually represented estimates of the location of parameters for each of the four facets alongside one another on the common logit scale for the construct. Thus, indicating the relative contribution of the sensory attributes on the construct of overall difference.

Several Rasch models were fitted to investigate data from the AR and DFC tests. The Rasch model equations for TIM and DFC Measures (DFCM) used in this study are outlined below. For both test results, one model includes all four facets, including the repetition facet for all three replicate datasets (TIM1 and DFCM1, …**Eq. 1** & …**Eq. 3**, respectively). The other three models exclude the repetition facet and have separate models fitted for each replicate dataset (TIM2 and DFCM2 for each replicate, **…Eq. 2**…**Eq. 4**, respectively).

The DFCM models have no attributes facet hence the absence of the $\delta_i$ (parameter)

$$(\text{TIM1})\ ln(P_{mnrik}/P_{mnrik-1}) = \beta_m - \theta_n - \rho_r - \delta_i - \tau_k \quad (1)$$

$$(\text{TIM2})\ ln(P_{mnik}/P_{mnik-1}) = \beta_m - \theta_n - \delta_i - \tau_k \quad (2)$$

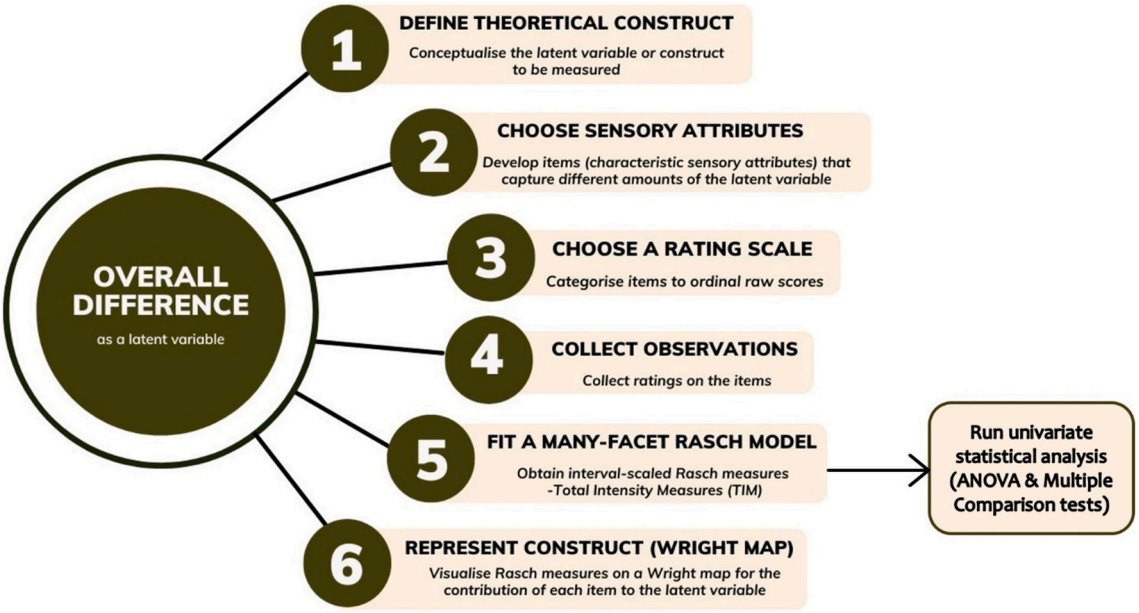


**Fig. 1.** Framework for the conceptualization of Overall Difference as a latent variable.

$$(\text{DFCM1})\ ln(P_{mnrk}/P_{mnrk-1}) = \beta_m - \theta_n - \rho_r - \tau_k \quad (3)$$

$$(\text{DFCM2})\ ln(P_{mnk}/P_{mnk-1}) = \beta_m - \theta_n - \tau_k \quad (4)$$

Where:

$P_{mnrik}$= probability that sample $(n)$ is rated $(k)$ for a sensory attribute $(i)$ by assessor $(m)$ in session $(r)$.

$P_{mnrik}$= probability that sample $(n)$ is rated $(k-1)$ for sensory attribute $(i)$ by assessor $(m)$ in session $(r)$.

$\beta_m$= degree of leniency or severity of assessor $(m)$ in rating attribute intensities.

$\theta_n$= degree of overall difference based on the total intensity measure for sample $(n)$.

$\rho_r$ = degree of difference between ratings of samples in a replicated session/repetition $(r)$

$\delta_i$ = degree of importance of a sensory attribute $(i)$ to the latent variable

$\tau_k$ = points on the latent variable continuum where the samples are equally likely to be rated between scale category $(k)$ and category $(k-1)$

### 2.5. Data analysis

All statistical analyses were conducted using R version 4.2.1. (R Core Team, 2022) while Rasch analysis was conducted using FACETS version 3.84.1 (Linacre, 2022a) and WINSTEPS® version 5.3.2 (Linacre, 2022b). DFC ratings and the attribute intensity ratings were each fitted to the MFRM, and then results from statistical analysis of the DFC raw scores, DFC Rasch measures, and Total Intensity Rasch measures were compared for discriminatory ability and diagnostic detail.

#### 2.5.1. Rasch analysis

*2.5.1.1. Fitting the Many-Facet Rasch Model (MFRM).* The MFRM considers the influence of multiple variables or explanatory factors (facets), and models all facets simultaneously on a common interval scale (i.e., the logit scale), with the log-odds of the raw score ratings as the outcome. The parameters for the facets were estimated using a Joint Maximum Likelihood Estimation (JMLE) method in the FACETS software (Linacre, 2022a), and location estimates for individual elements within each facet were plotted on a Wright map.

The Sample facet was non-centered, while the scale was adjusted for the other three facets (Assessor, Repetition, and Attribute) so that the mean of their parameters was centered at zero (0). This adjustment created a common reference point on the Wright map, around which the location of the samples relative to the Total Intensity Measure (TIM) was explored. Consequently, the location of the samples was adjusted by considering the severity of assessors, the intensity of attributes, and the intensity ratings in repeated sessions representing the Assessor, Attribute, and Repetition facets, respectively.

Following the recommendations of Linacre (2024a), all 4 facets were positively oriented on the Wright map such that higher Rasch measures generally mean higher ratings for all four facets. This is further illustrated in section **3.2. Representing the Construct of Overall Difference**.

*2.5.1.2. Global model fit.* The global fit of the data to the MFRM was examined. An acceptable fit (Linacre, 2022a) is when no more than 5 % of absolute standardized residuals is ≥2 and no more than 2 % is ≥3.

*2.5.1.3. Rating scale category diagnostics.* In Rasch analysis, scale category diagnostics are implemented to examine whether rating scales are functioning effectively. It reveals the operational use of the rating scale by assessors, and whether their interpretation of the scale categories aligns with the underlying construct measurement theory. This process is useful for guiding the revision of scoring materials to improve measurement and training procedures (Engelhard & Wind, 2018). When indicators for the proper functioning of rating scales are unmet, remedial actions to, as much as possible, extract the most reliable measures generally involve combining adjacent categories and sequentially renumbering the scale categories. Guidelines recommended by Bond et al. (2021); Eckes (2023); Ho (2019); Linacre (2002) for optimizing the functioning of rating scales have been summarized in ***Table A1***. Although some criteria for the proper functioning of rating scales were not met in this study, these issues do not impact the conclusions derived from the Rasch models. Further discussion on this point can be found in the results section- **3.1.2. Rating scale category diagnostics**.

*2.5.1.4. Facets model fit.* The method described in Ho (2019) was used to examine the fit of the assessor, sample, repetition, and attribute facets to the MFRM. Values between 0.5 and 1.5 are productive for

measurement, while values greater than 2.0 "*distort or degrade the measurement* system". Estimates with outfit mean-square values exceeding 2 were considered for removal only if they degraded the measure, as Linacre (2024c) reports that these values may be due to very few observations.

#### 2.5.2. Statistical analysis

To allow for adequate comparison between the DFC and the proposed Total Intensity Measure (TIM) approach for measuring overall differences between samples, data from the DFC raw scores, DFC Rasch measures, and Rasch measures of the combined attribute intensities (i.e. TIM) were fitted to statistical models separately, and all three results were compared.

Differences between the mean values of the 3 Jaffa cake samples were examined using parametric and non-parametric analysis of variance (ANOVA) models. The R packages *MASS* (Venables & Ripley, 2002), *car(*Fox & Weisberg, 2011*),* and *nortest* (Gross & Ligges, 2015) were used to fit parametric factorial ANOVA models and to conduct residual analysis. For non-parametric analyses, the *PMCMRplus* package (Pohlert, 2023) was used to conduct Friedman tests (Conover & Iman, 1981) and to conduct pairwise comparisons with a control using the *frdManyOneNemenyiTest* function (Hollander et al., 2014). A Bonferroni p-adjustment (Bonferroni, 1936) was specified to control for familywise error rates, and the alternative hypothesis was specified as "greater" for a one-tailed test.

## 3. Results

### 3.1. Fit of data to the many-facet Rasch model (MFRM)

Data from the AR and DFC tests were each fit to the MFRM and examined for evidence of their adequate fit to the Rasch model.

#### 3.1.1. Model fit statistics

Table 1 presents a summary of the Rasch model fit statistics for all 4 facets in the datasets (TIM1 and DFCM1). As described in **section 2.4. Defining the construct of overall difference as a composite of attribute intensity ratings**, additional Rasch models with 3 facets (excluding the repetition facet) were fitted for each replicate dataset from repeated sessions of the AR and DFC tests labeled as TIM2. Reps 1–3 and DFCM2.Reps 1–3, respectively.

The global model fit for the TIM1 showed the best fit to the Rasch model. Assessor fit improved in the models with the repetition facet included. This improvement likely occurred because including *Repetition* as an explanatory factor, and averaging ratings across the 3 sessions reduced inconsistent ratings within assessors. Assessor fit indices provide estimates of the consistency with which each assessor uses the rating scale categories across all facets (Eckes, 2023). While there is evidence that some assessors provided inconsistent ratings (as indicated by the Outfit mean-square in the Assessor facet), this misfit was not sufficient to degrade the measures. The reason is that all other facets demonstrated a 100 % fit to the Rasch model. Additionally, the impact of a few misfitting assessors on sample and attribute/item estimates is negligible (Wright & Linacre, 1994).

#### 3.1.2. Rating scale category diagnostics

Table 2 shows the scale category statistics for the DFC and Intensity rating scales. Mean Rasch estimates of the combined raw attribute ratings – Total Intensity Measures (TIM) were produced after fitting all 4 facets (assessor, sample, repetition, and attribute; **see 2.5.1. Rasch analysis**) to the MFRM. Both the 8-category intensity scale for TIM and the 7-category difference scale for the DFC measures failed to meet the criteria (see **Table A1**) for category precision in a proper functioning rating scale. This suggested that the affected categories were not used meaningfully by the assessors to distinguish between samples with respect to the respective underlying constructs.

For the TIM Intensity scale, the extreme category 7 *(Strongest imaginable oral sensation)* had less than 10 observations. While the DFC scale did not meet the criterion for a minimum advancing distance of 0.57 between Rasch-Andrich thresholds for a 7-category rating scale. Specifically, the threshold distances were 0.23 between categories 1 (barely detectable difference) and 2 (slight difference), and 0.26 between categories 5 (very large difference) and 6 (extremely different). However, while the criteria for category precision provide useful information for measurement inference, they are not essential. Therefore, revising the rating scales used in this study was deemed unnecessary, especially considering that reusing the scales for measurements across other samples with a similar context falls beyond the scope of this study. The focus was to explore the use of the MFRM approach for measuring overall differences between samples, rather than modifying measurement tools to enhance measurement procedures for Jaffa cakes.

### 3.2. Representing the construct of overall difference

The components of the Many-Facet Wright maps (Fig. 2 and Fig. 3) used in this study are outlined below.

**Table 1**
Summary of Rasch model fit statistics for DFC and Total Intensity Measure (TIM) models.

| | | | | OUTFIT Mean-Square[1] | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Global fit[2] | | | Assessor | | Sample | Repetition | Attribute |
| *Criteria* | % StRes *≤5 % ≥ 2* | % StRes *≤1 % ≥ 3* | Total[3] | % Fit *0.5–1.5*[4] | % Misfit *>2.0*[5] | % Fit | % Fit | % Fit |
| **TIM1** | **4.6 (138)** | **0.3(9)** | **3015** | **82** | **5** | **100** | **100** | **100** |
| TIM2.Rep1 | 4.5 (45) | 0.4 (4) | 1005 | 69 | 2 | 100 | NA[6] | 100 |
| TIM2.Rep2 | 4.9 (45) | 0.2 (2) | 1005 | 67 | 8 | 100 | NA | 100 |
| TIM2.Rep3 | 4.3 (43) | 0.4 (4) | 1005 | 61 | 8 | 100 | NA | 100 |
| **DFCM1** | **2.8 (17)** | **0 (0)** | **603** | **65** | **6** | **100** | **100** | **100** |
| DFCM2.Rep1 | 3.5 (7) | 0.5(1) | 201 | 40 | 10 | 100 | NA | 100 |
| DFCM2.Rep2 | 3.5 (7) | 0(0) | 201 | 52 | 13 | 100 | NA | 100 |
| DFCM2.Rep3 | 4.5 (9) | 0 (0) | 201 | 35 | 13 | 100 | NA | 100 |

[1] Outlier-sensitive measure of unweighted mean squares indicating deviation of the estimates of the four facets from predictions of the Rasch model.
[2] Percentage (number of observations in brackets) of absolute standardized residuals (StRes).
[3] Total number of responses used for the estimation of the model parameters.
[4] Outfit Mean-square values between 0.5 and 1.5 are considered productive for measurement (Linacre, 2024c). The same criteria apply to the percentage fit for all facets.
[5] Outfit Mean-square values >2.0 may degrade the measurement (Linacre, 2024c).
[6] NA implies Not Applicable as the Rasch models per replicate did not have a Repetition facet.

**Table 2**
Summary of scale category statistics for AR intensity and DFC rating scales in Rasch models - TIM1 and DFCM1 (which have all four facets - assessors, samples, repetition, and attributes fitted).

| Scale | Scale Categories | | Frequency[1] | Average Measure[2] | | OUTFIT Mnsq[3] | Rasch-Andrich Threshold | |
|---|---|---|---|---|---|---|---|---|
| | | | | Observed | Expected | | Measure | Distance[4] |
| **INTENSITY** | | | | | | | | |
| *Rating Scale* | 0 | None | 148 (5) | −2.26 | −2.03 | 0.8 | | |
| 8-category | 1 | Barely detectable | 392 (13) | −1.60 | −1.61 | 1.0 | −2.81 | 0.97 |
| 01234567 | 2 | Weak | 641 (21) | −1.00 | −1.08 | 1.0 | −1.84 | 0.65 |
| | 3 | Moderate | 937 (31) | −0.54 | −0.55 | 1.0 | −1.19 | 1.35 |
| | 4 | Strong | 583 (19) | −0.13 | −0.1 | 1.1 | 0.16 | 0.82 |
| | 5 | Very strong | 239 (8) | 0.23 | 0.25 | 1.0 | 0.98 | 0.65 |
| | 6 | Extremely strong | 69 (2) | 0.44 | 0.52 | 1.1 | 1.63 | 1.44 |
| | 7 | Strongest imaginable oral sensation | 6 (0)* | 0.88 | 0.73 | 0.9 | 3.07 | |
| **DFC** | | | | | | | | |
| *Rating Scale* | 0 | No difference | 69 (11) | −1.43 | −1.44 | 1.1 | | |
| 7-category | 1 | Barely detectable difference | 131 (22) | −0.82 | −0.83 | 1.1 | −1.71 | 1.00 |
| 0123456 | 2 | Slight difference | 135 (22) | −0.57 | −0.54 | 0.9 | −0.71 | 0.23* |
| | 3 | Moderate difference | 146 (24) | −0.26 | −0.26 | 1.0 | −0.48 | 0.97 |
| | 4 | Large difference | 79 (13) | 0.00 | 0.01 | 1.0 | 0.49 | 0.59 |
| | 5 | Very large difference | 31 (5) | 0.35 | 0.27 | 0.8 | 1.08 | 0.26* |
| | 6 | Extremely different | 12 (2) | 0.45 | 0.50 | 1.0 | 1.34 | |

[1] Total count (percentage distribution in brackets) of observations used in each scale category.
[2] Observed average measure (in log odds unit or logits), and expected average measure if data fits the Rasch model.
[3] OUTFIT Mean square refers to the outlier-sensitive measure of unweighted mean squares and indicates the deviation of responses from predictions of the Rasch model.
[4] Absolute difference between Rasch-Andrich threshold measures (in logits) of two adjacent scale categories. Where minimum distance for 8, 7 and 3 category scale = 0.51, 0.57 and 1.4 respectively; maximum difference = 5.0.
*Indicates criteria (Table A1) is unmet.

1. **First column – common logit scale**: shows the measure estimates for the four fitted facets in log-odds unit. The mean of measures in a logit scale is zero (0).
2. **Second column – assessor severity spread**: indicates the variation in severity among assessors.
3. **Third column – sample location**: shows the location of the samples along the construct based on average attribute intensity ratings (for TIM) or DFC ratings.
4. **Fourth column – variation in repeated ratings**: differences in ratings across the three repeated sessions
5. **Fifth column – attribute / item contribution**: represents the relative contribution of attributes/items (for TIM) to the overall construct; not applicable to the DFC Wright map.
6. **Rightmost column – rating scale**: displays the Rasch-half point thresholds as dotted horizontal lines (————). These thresholds mark the end of a category's interval (Linacre, 2024b), and indicate the point where the chance of a sample receiving a higher rating starts to exceed the chance of being rated on the lower adjacent category (Myford & Wolfe, 2003).

Note that scale category numbers in parenthesis (i.e. 0-None and 7-Strongest imaginable oral sensation) identify extreme ends of the scale. In Rasch measurements, the latent variable is conceptualized as "*infinitely long*" (Linacre, 2002). Consequently, the lowest and highest categories of the scale are "*infinitely wide*" to accommodate extreme responses by widening the scale as necessary.

As previously stated in **2.5.1. Rasch analysis**, all four facets were positively oriented on the Wright map. The (+) symbol indicates that higher measures on the logit scale correspond with the following. For the:

- *Assessor* Facet: assessors are more lenient and rated higher scores on the rating scale.
- *Sample* Facet: samples have higher Total Intensity Measure (TIM) or DFC measure (DFCM).
- *Repetition* Facet: samples were rated with higher intensity in a repeated session.
- *Attribute* Facet: attributes have higher average intensity ratings for each sample and have a higher contribution to the overall construct.

Fig. 2 and Fig. 3 represent the Many-Facet Wright maps for TIM1 and DFCM1 models respectively.

### 3.3. Total intensity measure (TIM1)

The TIM Wright map (Fig. 2) showed that assessors exhibited varying degrees of severity in their use of the intensity rating scale.

On average, attribute intensity ratings for the samples were below average (0) on the logit scale, and ratings across the three replicated sessions were consistent. Samples located higher on the scale were perceived to have higher intensities of the sensory attributes. Similarly, the *attributes* facet revealed the hierarchy of sensory attribute contributions to the sample differences. Orange flavor and sweetness were perceived as the most intense, demonstrating the highest contribution followed by cocoa flavor. Milky flavor and saltiness did not contribute as much to the overall difference. The *intensity scale* represents the category range within which the attributes were rated on average. Notably, the gaps between adjacent scale categories are not equidistant and tend to widen toward the extreme categories. The intensity scale revealed that all the samples were rated to have moderate differences in intensities for the combined attributes. Pairwise comparison tests against a control would determine the existence of significant overall differences between Brand A and Brand B compared to the Control, based on their Total Intensity Measures (TIM) from the logit scale.

### 3.4. DFC measure (DFCM1)

The Wright map for the DFCM1 (Fig. 3) revealed varying degrees of severity among assessors. Assessor 1011 consistently rated the samples using the lower end of the rating scale and emerged as the most severe assessor. For the *sample* facet, the difference from control for Brand A was rated higher than that for Brand B. While assessors rated Brand A as moderately different from CONTROL, the slight difference rating between Brand B and CONTROL was not significant. It was hypothesized that assessors may have considered differences perceived from other

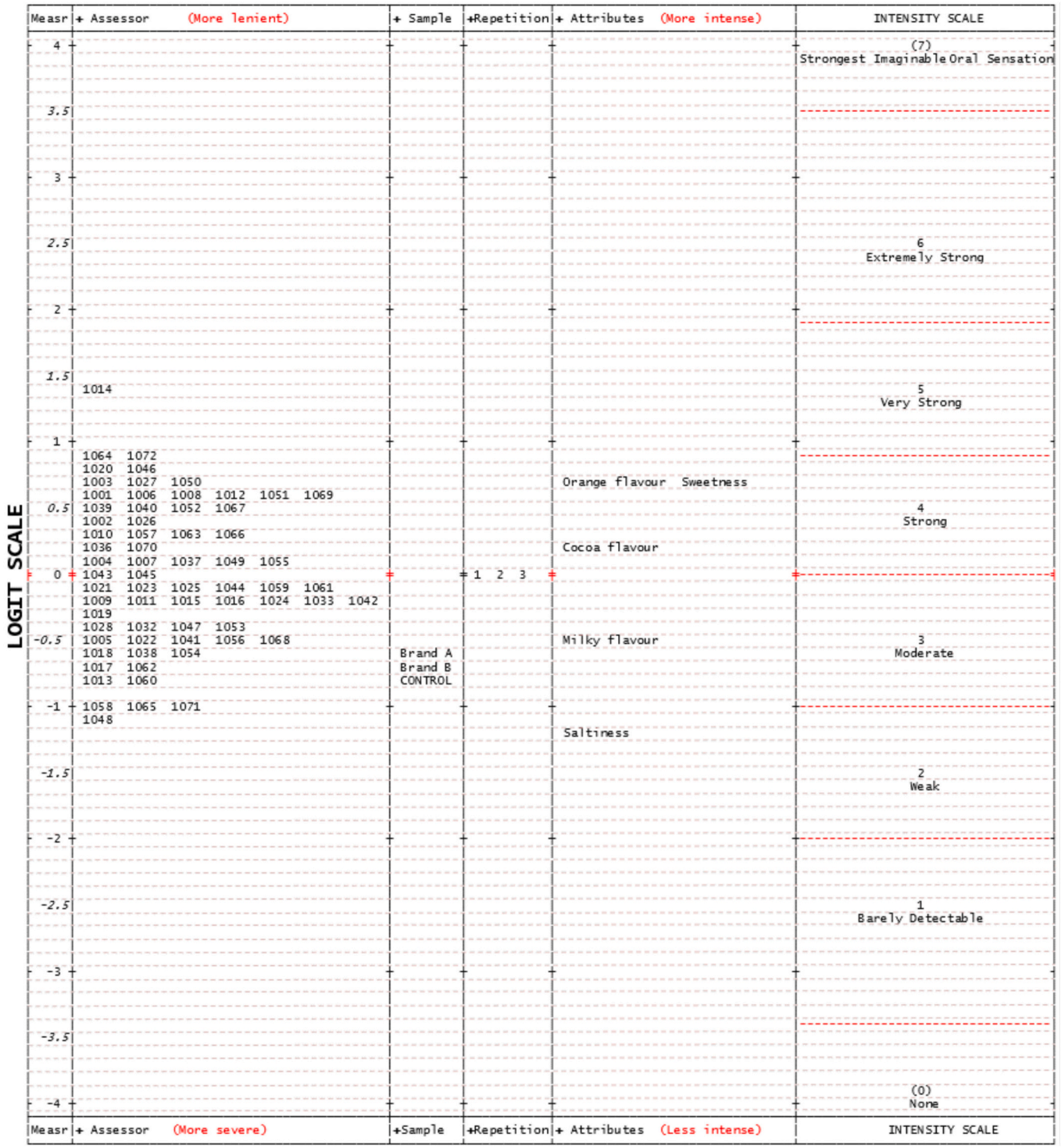


**Fig. 2.** Many-Facet Wright map for TIM1. The first column "Measr" represents Rasch model measures on the logit scale. The four facets are displayed from left to right: 1001–1072 represent unique assessor IDs for 67 assessors in the assessor facet; Brands A and B represent the test samples, and Control refers to the reference sample in the sample facet. Numbers 1–3 indicate replicate evaluations in the repetition facet, and attributes are listed in the attribute facet. The rightmost column illustrates the functioning of the AR intensity rating scale, with horizontal lines marking half-point thresholds, where the probability of a sample receiving a higher rating begins to exceed the likelihood of being rated in the lower adjacent category.

sensory modalities, such as appearance and texture, or other attributes that were not intended to be captured in the study. Efforts to maintain consistency across these attributes during sample selection may not have been entirely successful. Brand A had a slight difference in shape compared to the other samples (*Fig. B1*), which some assessors may have noticed. This is consistent with feedback from assessors after the study, who mentioned that they could easily identify Brand A due to their frequent consumption and familiarity with Jaffa cakes.

In the *repetition* facet, average DFC ratings increased in successive repeated sessions, with the third session showing the highest DFC ratings. This increase may be owing to assessors probably experiencing fatigue and some context bias from tasting numerous samples during the

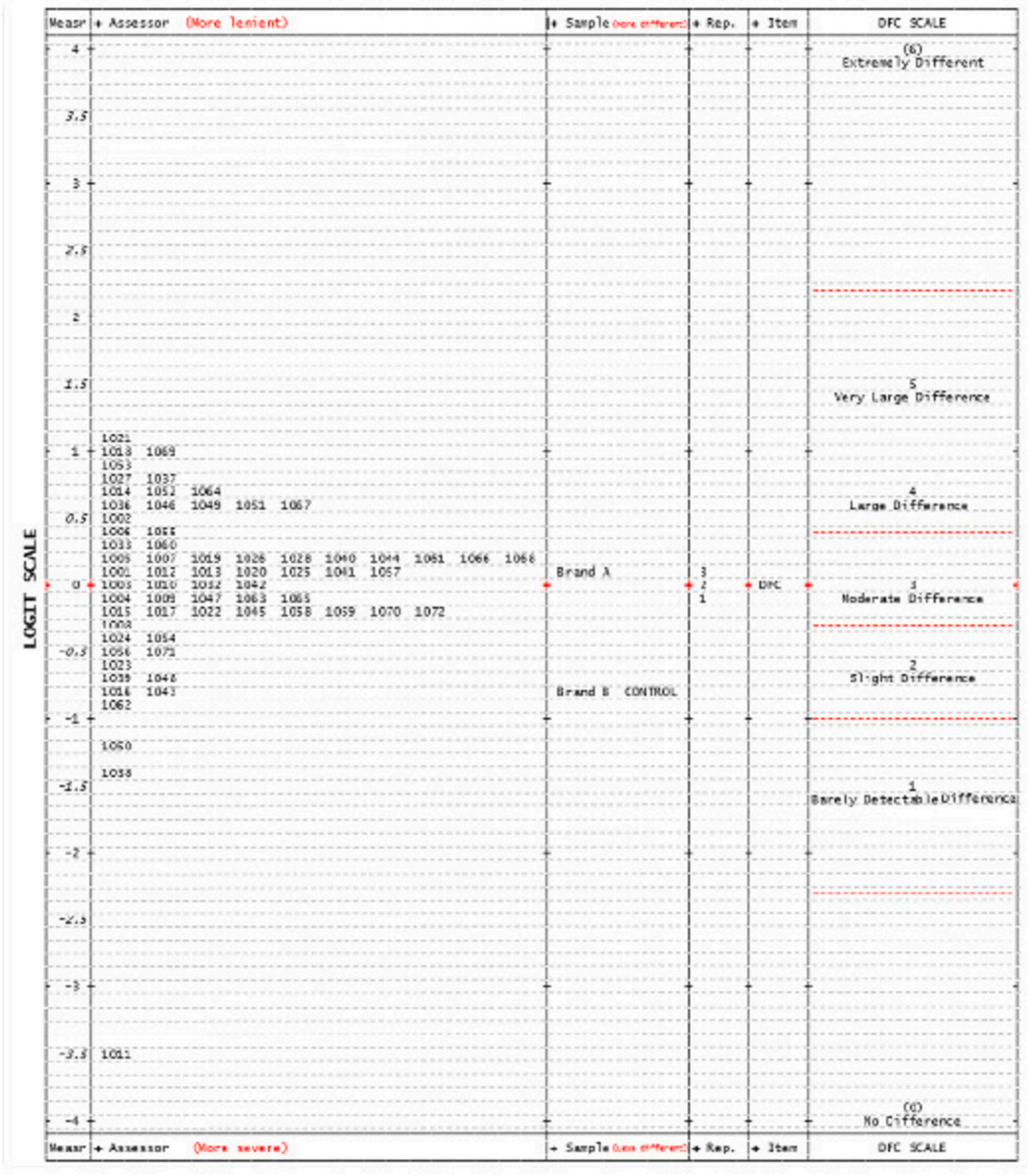


**Fig. 3.** Many-Facet Wright map for DFCM1. The first column "Measr" represents Rasch model measures on the logit scale. The four facets are displayed from left to right: 1001–1072 represent unique assessor IDs for the 67 assessors in the assessor facet; Brands A and B represent the test samples, and Control refers to the reference sample (R) in the sample facet. "Rep." denotes the repetition facet, with numbers 1–3 indicating replicate evaluations, and "item" refers to the single difference from control question use to evaluate the samples. The rightmost column illustrates the functioning of the difference rating scale for the DFC, with horizontal lines marking half-point thresholds, where the probability of a sample receiving a higher rating exceeds the likelihood of being rated in the lower adjacent category.

test. Like the TIM Wright map, the gaps between adjacent scale categories are not equidistant and tend to widen toward the extreme categories (the interval range for end categories (0) and (6) is not fully captured on the Wright map).

### 3.5. Comparing the overall difference between samples

Table 3 summarizes the statistical test results for the TIM and DFCM Rasch models, as well as the datasets from DFC RAW scores, along with their replicate datasets. Strata and Reliability values from Rasch separation statistics are also presented. *Strata* refers to the number of statistically distinct groups distinguishable by the respondents in a measurement instrument (Myford & Wolfe, 2003; Wright & Masters, 2002). A Strata of 1 indicates that the instrument cannot reliably distinguish between different levels of the latent variable. 2 Strata shows a distinction between high and low levels only. 3 Strata indicate low, medium, and high levels of a latent variable while 4 or more Strata signify that the instrument can distinguish between 4 or more distinct groups. Low Strata statistics may suggest a need to add more discriminative items or refine existing ones to capture more of the latent variable. On the other hand, the *Reliability* index indicates whether differences found between the samples are due to measurement error. A Reliability value <0.50 suggests that differences between measures are primarily due to measurement error *(*Wright & Masters, 1982*)*. This could be due to a lack of variation in responses or insufficient items.

All datasets for the DFC Rasch measures revealed statistically significant distinct levels for the samples, evidenced by Strata values greater than 4 and Reliability values close to 1.0. This suggested that there was a statistically significant difference between at least one of the samples and the control sample based on their DFC ratings in all three repeated sessions.

Strata for the samples in TIM varied between repeated sessions. For the first two replicate sessions (TIM2.Rep1 and TIM2.Rep2), Strata values were less than 2 with Reliability values less than 0.5 suggesting that the assessors could not distinguish between the samples based on the average attribute intensity ratings. TIM2.Rep3 revealed a distinction between high and low levels of intensities for the sample with a Strata value of 2 and a Reliability statistic greater than 0.5. However, upon averaging across the three repeated sessions, a Strata value of 3 for TIM1 indicated three statistically distinct levels for the samples, supported by a Reliability value closer to 1.0. This suggests that averaging across replicated sessions helped reduce inconsistencies in assessor ratings, and that the Rasch model accounted for variations in the severity of these averaged ratings, thereby improving the discriminatory ability of the measurement.

Parametric two-way ANOVA tests also revealed the existence of significantly different samples in all the datasets**.** However, they all failed to meet the assumptions for parametric ANOVA models upon conducting residual analysis. Non-normality was detected in both the TIM and DFCM estimates, and Breusch-Pagan tests revealed that residuals for all datasets (DFC RAW Scores, DFCM models, and TIM models) were heteroscedastic.

Since all three datasets violated ANOVA assumptions, non-parametric rank sums were used for mean comparisons. The Friedman tests (with $p < 0.01$) indicated significant differences between the samples, corroborating the findings from the parametric two-way ANOVA and Rasch separation statistics. However, since the Friedman test is designed for unreplicated data, the replicated measures for all the samples were averaged across all assessors before conducting the Friedman tests.

Pairwise comparisons using a Nemenyi-Wilcoxon-Wilcox-Miller many-to-one test for a two-way balanced complete block design showed that for both the DFC RAW and DFCM1, only Brand A was significantly different from the CONTROL. TIM results revealed that both Brands A and B were significantly different from the CONTROL suggesting that perhaps requiring assessors to focus on specific attributes revealed differences between perceived intensities of the attributes for all the samples.

In relation to Rasch separation statistics, Strata values were higher for the DFC Rasch models compared to TIM. Assessors could identify more distinct levels of difference for the samples by rating the overall difference from the control. These ratings could have been influenced by perceived differences other than the taste of the samples. As previously discussed in **3.2: DFC Measure (DFCM1)**, the perceived difference in non-taste attributes and familiarity with Brand A may have influenced assessors' DFC ratings, despite attempts to eliminate this effect. In

**Table 3**
Comparison of Sample facet summary statistics for all TIM and DFCM Rasch models (with Repetition facet - TIM1/DFCM1 and without Repetition facet - TIM2/DFCM2) and RAW DFC scores (individual replicates and averaged).

| Test/ Dataset[1,2] | TIM Models | | | | DFCM Models | | | | DFC RAW Scores | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TIM2. Rep1 | TIM2. Rep2 | TIM2. Rep3 | TIM1 | DFCM2. Rep1 | DFCM2. Rep2 | DFCM2. Rep3 | DFCM1 | Rep1 | Rep2 | Rep3 | Averaged Reps |
| **Rasch Separation Statistic** | | | | | | | | | | | | |
| Reliability $_{Sample}$ | 0.45 | 0.35 | 0.68 | **0.83** | 0.94 | 0.92 | 0.97 | **0.98** | | | | |
| Strata $_{Sample}$ | 1.53 | 1.31 | 2.27 | **3.31** | 5.40 | 5.01 | 7.86 | **8.78** | | | | |
| **ANOVA Residual Analysis (*P*-values)** | | | | | | | | | | | | |
| **Normality** | | | | | | | | | | | | |
| Shapiro-Wilks | <0.001 | <0.001 | <0.001 | **<0.001** | <0.001 | <0.001 | <0.001 | **<0.001** | 0.043 | 0.071 | 0.311 | **0.290** |
| **Outlier Test** | | | | | | | | | | | | |
| Bonferroni | 0.033 | NA | NA | **NA** | NA | 0.243 | <0.001 | **NA** | 0.026 | 0.683 | 0.319 | **0.034** |
| **Constancy of Error Variance** | | | | | | | | | | | | |
| Breusch-Pagan | <0.001 | 0.006 | <0.001 | **0.081** | 0.070 | <0.001 | <0.001 | **0.271** | <0.001 | 0.011 | 0.002 | **<0.001** |
| **Friedman Test** | | | | | | | | | | | | |
| $X^2$ | 134*** | 134*** | 134*** | **134***** | 134*** | 134*** | 134*** | **134***** | 20.39*** | 14.21*** | 45.80*** | **46.72***** |
| **Nemenyi Many to One Test (Pairwise Comparisons)** | | | | | | | | | | | | |
| ***Mean differences*** | | | | | | | | | | | | |
| Control-Brand A | −0.19*** | −0.08*** | −0.23*** | **−0.19***** | −1.13*** | −0.92*** | −1.43*** | **−0.82***** | −0.94*** | −1.01*** | −1.39*** | **−1.11***** |
| Control-Brand B | −0.07*** | −0.07*** | −0.08*** | **−0.07***** | 0.02 | −0.01*** | 0.20 | **0.05** | 0.01 | −0.01 | 0.18 | **0.06** |

[1] *P*-value levels of significance: <0.001***, <0.01**, <0.05*; measures with no superscript symbols >0.05.
[2] NA implies Not applicable as no outliers were found.

comparison, low Strata values for TIM suggest that the range of taste attributes selected to capture the latent variable of overall difference could be refined to be more discriminative. Perhaps a different set of taste attributes or even the inclusion of other sensory modalities may be helpful in distinguishing better between the samples based on combined ratings.

## 4. Discussion

### 4.1. Measuring overall difference with the MFRM

The Many-Facet Rasch Model (MFRM) has been employed to generate estimates from a combination of sensory attributes (the Total Intensity Measures - TIM) that measure the latent trait of overall difference between samples. This TIM method was then combined with pairwise comparison against a control, enabling the estimation of overall product differences relative to a reference sample. The results demonstrated that this approach could be equally as effective as the Difference-from-Control (DFC) method in comparing differences against a control product. It yielded valuable quantitative data capturing nuanced differences between products by quantifying and providing a hierarchy of sensory attribute contributions to perceived differences. Additionally, it allows for comparisons to be made either between individual test products or between the test products and a control using the appropriate statistical test. In contrast, the DFC only allows for comparisons with a control and never between individual test samples (Rogers, 2017). The control sample can be predetermined during conceptualization or retrospectively selected, and an action standard can be established to guide decisions regarding the implementation of product changes.

### 4.2. Rasch-transformed rating scales

Rasch models allow linear measurement of latent variables using ordinal response data. Resulting interval-scaled measures expressed in log-odds units (logits) enhance the interpretability of ordinal responses from labeled category scales. Wright maps (Fig. 2 and Fig. 3) visualize individual scale categories as threshold ranges, indicating the transition points between rating categories. Slightly unequal distances between adjacent scale categories, which widen toward the extreme ends of the scale, are characteristic of category-ratio scales like the Labeled Magnitude Scale (LMS) (Green et al., 1993), the Borg Scale (Borg, 1982), and the generalized Labeled Magnitude Scale (gLMS) (Bartoshuk et al., 2005). Rather than requiring assessors to learn complex category-ratio scales to choose a single point on the scale describing their perception, simpler categorical-labeled ordinal scales can be used. The MFRM can then produce similar interval measures. This approach is likely to reduce variability in assessor responses (Ho, 2019).

Rating scale category statistics and graphs from Rasch analysis offer insights into how assessors interpret and use individual rating scale categories in an experiment. An empirical investigation, following established guidelines (Table A1), reveals deviations in the interpretation and operational use of the rating scale from the Rasch model's expectations for the conceptualized latent variable being measured (Engelhard & Wind, 2018). This information is valuable for improving rating scales by eliminating redundant categories or filling gaps in scale categories. It also informs revisions to panel training and measurement procedures. From a manufacturer's perspective, a Rasch approach can help design long-term sensory quality programs for specific products with an effective rating scale.

Based on the rating scale category diagnostics (as shown in Table 2) and the Wright maps, it becomes evident that the assessors rarely utilized the extreme ends of the scale when rating the Jaffa cake samples. If the intention is to apply these scales across various tests measuring overall differences between Jaffa cakes, the insights gained from the model's identification of redundancies in the scale can inform improvements to enhance the functionality of the rating scales.

### 4.3. Monitoring assessor performance

The benefits of the MFRMs in "*rater-mediated assessments*" (Eckes, 2023; Engelhard & Wind, 2018) have been reported by several researchers (Bond et al., 2021; Boone et al., 2014; Engelhard, 2013; Engelhard & Wind, 2018; Linacre, 1994; Myford & Wolfe, 2003). In sensory measurements, the expectation that the human assessors function as a unanimous instrument is unrealistic. Individual variability always introduces complexities, regardless of how much training assessors receive (Bartoshuk et al., 2005; Meilgaard et al., 2016; Næs et al., 2010; Sipos et al., 2021; Stone et al., 2012).

With Rasch measurement, the goal is consistency within individual assessors in terms of severity level and the understanding of the rating scale, rather than unanimous panel ratings (Linacre, 1994). Quality control parameters of the model, such as outlier sensitive measures (OUTFIT mean-square), can identify unwanted idiosyncrasies in individual ratings. Additionally, questionable individual rating patterns can be provided as feedback to assessors, encouraging improvement (Findlay et al., 2007). Consequently, costs associated with repeated assessor training can be reduced.

### 4.4. Why use a TIM approach?

This study highlights the strengths of the TIM approach in sensory quality control. Its flexibility makes it well-suited for targeted evaluations, such as identifying quality issues, benchmarking against standards or competitors, and exploring differences in specific attributes across samples. TIM can work as both an overall difference test and a tool for analyzing how individual attributes contribute to those differences within a single analysis.

By combining targeted attribute assessments with holistic difference estimation, TIM addresses key limitations of traditional methods while leveraging their strengths. Unlike the DFC method, which restricts comparisons to a control sample, it allows for comparisons between individual test products and for comparisons against a control. As it relies on attribute rating tests, multiple attributes can be simultaneously evaluated, requiring fewer samples than the DFC and fewer tests than attribute-specific methods like paired comparisons or n-AFC tests, which assess one attribute at a time. This makes TIM a practical and cost-effective approach, especially in resource-constrained settings.

The ability of the TIM approach to consolidate attribute intensity ratings into a single measure of overall difference using the MFRM allows analysts to estimate overall product differences based on attributes of interest, capturing meaningful variations between products while filtering out non-critical variations that might otherwise distract from the analysis. The Wright map enhances this by visually representing the relative contribution of these attributes in driving observed differences, making it easier to identify specific issues and take targeted action.

With the MFRM, multiple variables such as assessors, attributes, products, replicates, order effects, and other factors can be integrated into a single, unified model. This eliminates the need for separate statistical analyses, like multiple ANOVAs to assess individual assessor performance or multivariate techniques like PCAs to explore product relationships. Instead, TIM operates on a unidimensional construct, assuming that all factors can be measured on the same Rasch logit scale. The built-in Rasch quality control metrics allow for the simultaneous monitoring of each variable, ensuring that they align with the model expectations.

In Rasch analysis, each assessor's ratings are treated individually, and their level of severity (the tendency to rate higher or lower compared to others) is accounted for in the model. By simultaneously estimating both the difficulty of the items (i.e., attribute intensities) and assessor severity, the model allows for more accurate comparisons across assessors with different standards, removing the need for

extensive training on complex rating scales. Assessors can rate according to their own consistent standards, as long as these are consistent across evaluations. Additionally, the model converts ordinal data into interval-scale data, enabling the effective use of categorical rating scales for rating intensity, provided assessors are trained to understand where attribute intensities fall on the scale for the specific products being evaluated.

While TIM offers significant advantages, it is not meant to replace traditional sensory methods. Instead, it serves as a complementary tool, providing unique benefits when deeper, more focused insights are needed. Although assessors still require some training and product sensory specifications need to be defined, TIM has the potential to streamline processes by monitoring attribute contributions to product differences, assessor performance, consistency in ratings across evaluations, and the functionality of rating scale categories within a single analysis. This delivers actionable data for both product diagnostics and assessor selection and training, making TIM a valuable addition to sensory quality control methods.

### 4.5. Limitations of the study

In this study, the TIM approach was used to assess overall differences between samples and a prescribed control based on a combination of attribute intensity ratings, addressing the limitations of assessor fatigue and the resource-intensive requirements of the DFC. However, since the comparison was made between the overall difference results from the DFC and those from Rasch-combined taste attribute ratings, the selection of test products and attributes did not fully account for differences that might have been perceivable during the DFC test.

As an overall difference test, the DFC allows assessors to either differentiate samples based on the most prominent perceived attribute difference, or average across all perceived attributes before making a distinction. As a result, some assessors may have considered additional sensory aspects beyond taste attributes in rating the Jaffa cake samples, yielding differing results to that of the TIM. The former was the case for Brand A, influenced by assessors' familiarity and possibly appearance and being rated as significantly different from the control, while Brand B was not.

In contrast, the AR test focused solely on taste attributes, leaving potential variations in other sensory characteristics unaccounted for. Consequently, the total intensity measure (TIM) was estimated based on only these taste attributes and overall intensity differences across all samples were rated as moderate on the latent variable. This narrow focus may have increased the risk of a *"Type 1 error"* in the TIM approach, identifying differences that might not fully represent overall product perception, or *"Type 2 error"* for the DFC whereby some assessors may have missed meaningful differences in the samples because they used a cognitive strategy of focusing on the most prominent attribute difference from the reference sample "R" which may not have been included in the AR test. Incorporating a broader range of attributes or integrating other sensory modalities could have reduced these potential errors, improved measurement accuracy, and strengthened the comparison between TIM and the DFC.

However, the TIM approach is advantageous when it is necessary to focus assessors on pre-selected, relevant attributes - those most likely to vary due to process changes, ingredient modifications, or product life-cycle stages, while reducing the risk of assessors basing their ratings on irrelevant or non-critical attributes. To enhance future comparisons of the TIM and DFC approaches, it is recommended that all attributes that would be perceivable in an overall assessment of the test samples, as done in the DFC, be included in the Attribute Rating (AR) test to ensure a more comprehensive evaluation. This can be achieved by conducting preliminary sensory tests to identify and guide the choice of attributes, ensuring a more robust comparison between the two approaches.

### 4.6. Conclusion

As always, the objective of a sensory test should inform the choice of the testing method. If the goal is to quantify how sensory attributes contribute to differences between products, the Total Intensity Measure (TIM) approach proves efficient. Statistical tests for pairwise comparisons with a control enable the measurement of overall sample differences using Total Intensity Measures (TIM), akin to the approach employed in the DFC.

However, *"rater-mediated assessments"* like sensory evaluations, should be iterative as noted by Engelhard and Wind (2018). Empirical results should inform future experimental designs, rating scale development, and assessor training procedures. The location of facet parameters on a Wright map depends on the assessors' intensity ratings for the specific attributes being assessed in a product. Consequently, whether more lenient or more severe in assigning ratings, individual assessors within a panel must maintain consistency in their ratings, effectively acting as individual experts. Rasch analysis addresses individual variations in the severity of ratings, provided there is a good fit between the model and the data. From a quality control standpoint, the TIM approach is ideal for establishing and enhancing sensory quality programs, especially when specific attributes of interest are well-defined, and quantitative assessments of these attribute contributions to perceived differences would guide product development and optimization decisions. Sensory lexicons from previous descriptive analyses may aid in identifying those relevant attributes.

This study demonstrated that a Rasch approach to measuring overall difference using a combination of sensory characteristics (TIM) can serve as an equally effective alternative to the DFC, with added benefits of evaluating targeted sensory attributes and revealing the relative importance of each attribute on the product differences. Further studies will explore the potential of MFRM in examining assessor performance during sensory evaluation.

**Funding sources**

The work reported here is part of N. C. Ariakpomu's doctoral research, funded by the Commonwealth Scholarship Commission and the Foreign, Commonwealth and Development Office in the UK. We are grateful for their support. All views expressed here are those of the authors, not the funding body.

**Data statement**

The data used for this research is available from the University of Leeds RADAR *and referenced here as (*Ariakpomu et al., 2024*)*.



**Dataset Reference:** Ariakpomu, N., Ho, P., & Holmes, M. (2024). Sensory attribute and difference from control ratings of Jaffa cakes.doi: https://doi.org/10.5518/1484.

**Ethical Statement**

Ethical approval for the involvement of human subjects in this study was granted by the University of Leeds Faculty Research Ethics Committee (AREA FREC), Reference number AREA FREC 2023–0433-496, dated 04/13/2023.

## CRediT authorship contribution statement

**Nnenna C. Ariakpomu:** Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Melvin J. Holmes:** Writing – review & editing, Supervision, Resources, Conceptualization. **Peter Ho:** Writing – review & editing, Supervision, Resources, Methodology, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We are grateful to all those who participated in the sensory evaluation studies, as well as the laboratory and school administrative staff who assisted with the logistics involved in planning the experiments.

## Appendix A. Appendix

**Table A1**
Guidelines for assessing effective rating scale[1].

| Criteria | Description | Implication |
|---|---|---|
| **Item Polarity** | Scales should be oriented in the same way as the latent variable, so that higher ratings imply more of the latent variable. This is essential for the description of the samples and for measure stability, measure accuracy, and inference about the different samples. | The Point Biserial Measure (PT measure) correlation should not be negative as this means that items do not align with the theoretical expectation of how the latent variable should be measured. Affected items should be rescored. |
| **Category Frequency** | There should be at least 10 observations in each scale category. This is essential for measure stability. | Category thresholds may be estimated poorly making it difficult for categories to describe distinct locations on the latent variable. Combine adjacent categories and renumber the categories in sequence to resolve. |
| **Category Frequency Distribution** | Frequency distribution of scale categories should be unimodal and tend toward a uniform distribution. | Intermittent low-frequency categories within the distribution may indicate irregular scale usage and the presence of redundant categories. To resolve, combine adjacent categories and renumber the scale categories in sequence. |
| **Observed Average Measures** | Computed as the average of the combined measure statistics of all the facets involved in producing scale category ratings. It should monotonically increase as the scale categories advance. Essential for measure accuracy and description of the samples. | Higher average measures will indicate ratings in higher scale categories and vice versa. A disordered category should be combined with adjacent categories. |
| **Category model fit** | Scale category outfit mean-squares indicate the deviation of average measure from the expected measures if data fit the Rasch model. Essential for measure accuracy. | Category outfit mean-square statistics with values above 2.0 indicate that the category has been used in a different context than is expected. |
| **Ordering of category thresholds** | Rasch-Andrich thresholds should advance monotonically up the scale categories. Graphical probability curves produced should have distinct peaks, resembling a range of hills. | As scale categories increase along the latent variable, each category, in turn, should be the most probable choice. Disordered thresholds may indicate that a category has been skipped as one advances along the variable or that the category has a very low frequency. |
| **Distance between category thresholds** | The minimum distance between Rasch-Andrich thresholds is calculated[2] as 1.4, 1.1, 0.81, 0.70, 0.57, 0.51, and 0.45 logits for 3, 4, 5, 6, 7, 8, and 9 category scales, respectively. The increase between thresholds should not exceed 5.0 logits. | Too close categories may be less distinctive than intended, while categories too far apart represent performance that is much wider than intended and introduces gaps in the variable leading to loss of information. |

[1] Bond et al. (2021); Eckes (2023); Ho (2019); Linacre (2002)
[2] Central distance $=ln(x/(m-x+1)$. For $x=1, \ldots m$, where $m=n$-1 for a $n$-category scale

## Appendix B. Sample Appearance

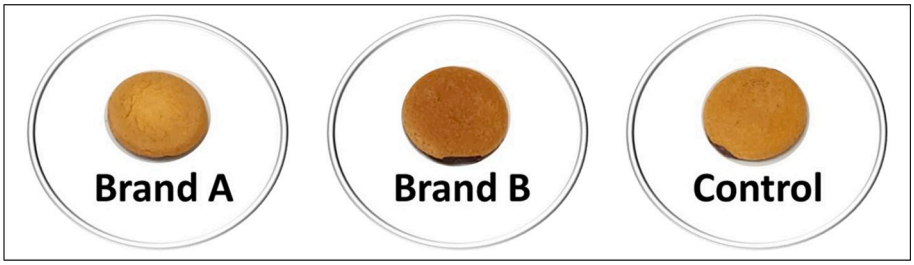


**Fig. B1.** Photo of Jaffa cake samples showing variation in appearance. Brand A exhibits greater variation in shape compared to Brands B and C. Note that the labels "Brand A," "Brand B," and "Control" are used here only to mask the actual brand names. During the study, all samples were labeled with random 3-digit codes.

## Data availability

The data for this study is available upon request through the University of Leeds Restricted Access Data Repository (RADAR) at https://doi.org/10.5518/1484

## References


Alvarez, P., & Blanco, M. A. (2000). Reliability of the sensory analysis data of a panel of tasters. *Journal of the Science of Food and Agriculture, 80*, 409–418. https://doi.org/10.1002/1097-0010(200002)80:3%3C409::AID-JSFA551%3E3.0.CO;2-T

Ares, G., Bruzzone, F., Vidal, L., Cadena, R. S., Giménez, A., Pineau, B., … Jaeger, S. R. (2014). Evaluation of a rating-based variant of check-all-that-apply questions: Rate-all-that-apply (RATA). *Food Quality and Preference, 36*, 87–95. https://doi.org/10.1016/j.foodqual.2014.03.006

Ariakpomu, N., Ho, P., & Holmes, M. (2024). *Sensory attribute and difference from control ratings of Jaffa cakes*. https://doi.org/10.5518/1484

Bartoshuk, L. M., Fast, K., & Snyder, D. J. (2005). Differences in Our Sensory Worlds: Invalid Comparisons With Labeled Scales. *Current Directions in Psychological Science, 14*(3), 122–125. https://doi.org/10.1111/j.0963-7214.2005.00346.x

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences (Fourth ed.). Routeledge*.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8*, 3–62.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE life. Science Education, 15*(4). https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. *Springer*. https://doi.org/10.1007/978-94-007-6857-4

Borg, E., & Kaijser, L. (2006). A comparison between three rating scales for perceived exertion and two different work tests. *Scandinavian Journal of Medicine & Science in Sports, 16*(1), 57–69. https://doi. https://doi.org/10.1111/j.1600-0838.2005.00448.x.

Borg, G. A. (1982). Psychophysical bases of perceived exertion. *Medicine and Science in Sports and Exercise, 14*(5), 377–381 (PMID: 7154893).

Compusense. (2020). Quality Assurance with Difference from Control Testing [White paper]. Retrieved 04 March 2023, from https://compusense.com/wp-content/uploads/2020/03/Difference_from_Control_Testing_White_Paper.pdf.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35*(3), 124–129. https://doi.org/10.2307/2683975

Costell, E. (2002). A comparison of sensory methods in quality control. *Food Quality and Preference, 13*(6), 341–353. https://doi.org/10.1016/S0950-3293(02)00020-4

Eckes, T. (2023). *Introduction to many-facet Rasch measurement*. Peter Lang.

Engelhard, G. (2013). *Invariant measurements: Using Rasch models in the social, behavioural, and health sciences*. Routledge.

Engelhard, G. J., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Faye, P., Courcoux, P., Giboreau, A., & Qannari, E. M. (2013). Assessing and taking into account the subjects' experience and knowledge in consumer studies. Application to the free sorting of wine glasses. *Food Quality and Preference, 28*(1), 317–327. https://doi.org/10.1016/j.foodqual.2012.09.001

Findlay, C. J., Castura, J. C., & Lesschaeve, I. (2007). Feedback calibration: A training method for descriptive panels. *Food Quality and Preference, 18*, 321–328. https://doi.org/10.1016/j.foodqual.2006.02.007

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression (2nd ed.). Sage*.

Gill, V., Ho, P., Ariakpomu, N., & Holmes, M. (2024). *Sensory attribute ratings of chocolate spreads.*. https://doi.org/10.5518/1483

Green, B. G., Dalton, P., Cowart, B., Shaffer, G., Rankin, K., & Higgins, J. (1996). Evaluating the 'labeled magnitude scale' for measuring sensations of taste and smell. *Chemical Senses, 21*(3), 323–334. https://doi.org/10.1093/chemse/21.3.323

Green, B. G., Shaffer, G. S., & Gilmore, M. M. (1993). Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chemical Senses, 18*(6), 683–702.

Gross, J., & Ligges, U. (2015). Nortest: Tests for normality. In *R package version* Version 1.4. https://cran.r-project.org/package=nortest.

Higgins, M. J., & Hayes, J. E. (2020). Discrimination of isointense bitter stimuli in a beer model system. *Nutrients, 12*(6), 1560. https://doi.org/10.3390/nu12061560

Ho, P. (2019). A new approach to measuring overall liking with the many-facet Rasch model. *Food Quality and Preference, 74*, 100–111. https://doi.org/10.1016/j.foodqual.2019.01.015

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric statistical methods (3rd ed. ed.). Wiley*.

Lawless, H. T., & Heymann, H. (2010). *Sensory evaluation of foods: Principles and practices (second, Ed.). Springer*.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2022a). *Facets computer program for many-facets Rasch measurement, version 3.84.1. In Winsteps.com*.

Linacre, J. M. (2022b). *Winsteps® Rasch measurement computer program (version 5.3.2). In Winsteps.com*.

Linacre, J. M. (2024a). Positively-oriented facet = 1. [online facet manual]. Retrieved 14 March 2024 from. https://www.winsteps.com/facetman64/positive.htm.

Linacre, J. M. (2024b). *Rating scale conceptualization: Andrich, Thurstonian, half-point thresholds [Online]. Retrieved 17th March 2024 from https://www.winsteps.com/winman/ratingscale.htm#:~:text=The%20Rasch%2Dhalf%2Dpoint%20thresholds,are%20shown%20in%20Table%2012.5.&text=3.,Cumulative%20Probabilities%22%20are%20of%20interest*.

Linacre, J. M. (2024c). *What do Infit and outfit, mean-square and standardized mean? [online]. Retrieved 14th march 2024 from https://www.rasch.org/rmt/rmt162f.htm*.

Meilgaard, M. C., Civille, G. V., & Carr, B. T. (2016). *Sensory evaluation techniques* (5th ed.). CRC Press.

Muñoz, A. M., Civille, G. V., & Carr, B. T. (1992). *Sensory evaluation in quality control*. Van Nostrand Reinhold.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4*(4), 386–422.

Næs, T., Brockhoff, P., & Tomic, O. (2010). Statistics for sensory and consumer. *Science*. https://doi.org/10.1002/9780470669181

Pohlert, T. (2023). Pmcmrplus: Calculate pairwise multiple comparisons of mean rank sums extended. In *R package version* (Version 1.9.10). https://cran.r-project.org/web/packages/PMCMRplus/index.html.

R Core Team. (2022). R. A language and environment for statistical computing (version 4.2.1). In R foundation for statistical computing, Vienna Austria. https://www.R-project.org/.

Redjade Software Solutions, L. (2023). Redjade sensory software. In. https://redjade.net/sensory-testing-software/.

Reinbach, H. C., Giacalone, D., Ribeiro, L. M., Bredie, W. L. P., & Frøst, M. B. (2014). Comparison of three sensory profiling methods based on consumer perception: CATA, CATA with intensity and napping. *Food Quality and Preference, 32*, 160–166. https://doi.org/10.1016/j.foodqual.2013.02.004

Rogers, L. (2017). *Discrimination testing in sensory science: A practical handbook (First ed.). Woodhead Publishing*.

Sipos, L., Nyitrai, Á., Hitka, G., Friedrich, L. F., & Kókai, Z. (2021). Sensory panel performance evaluation—Comprehensive review of practical approaches. *Applied Sciences, 11*(24), 11977. https://doi.org/10.3390/app112411977

Stone, H., Bleibaum, R. N., & Thomas, H. A. (2012). *Sensory evaluation practices (fourth edition / Herbert Stone, Rebecca N. Bleibaum, heather a. Thomas. Ed.). Academic*.

Thompson, M. (2003). The application of Rasch scaling to wine judging. *International Education Journal, 4*(3), 201–223.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S (4th ed.). Springer*.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis. MESA press*.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions, 16*(3), 888. https://www.rasch.org/rmt/rmt163f.htm.