Application of Rasch Analysis in Sensory Difference Testing

Nnenna Cynthia Ariakpomu

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Food Science and Nutrition

July 2025

Intellectual Property Rights Statement

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 4: ***Measuring Overall Difference with the Many-Facet Rasch Model (MFRM): The Total Intensity Measure (TIM) Method*** is based on the jointly authored publication with the following details.

Publication

Ariakpomu, N. C., Holmes, M. J., & Ho, P. (2025). Measuring overall difference from a combination of attribute ratings with the many-facet Rasch model. Food Quality and Preference, 127, 105442. https://doi.org/10.1016/j.foodqual.2025.105442

Author Contributions

Ariakpomu, N. C.: conceptualisation, writing (original draft, revisions and refinement), methodology design, ethics application, data collection, analysis, visualisation, curation and publication in Research Data Leeds Repository, funding acquisition, project management and administration.

Holmes, M. J.: supervision, conceptualisation, and reviewing manuscript.

Ho, P.: supervision, conceptualisation, reviewing manuscript, methodology design, provision of Rasch analysis software, and data analysis.

Research Outputs and Professional Development

Published article

- **Ariakpomu, N. C.**, Holmes, M. J., & Ho, P. (2025). Measuring overall difference from a combination of attribute ratings with the many-facet Rasch model. Food Quality and Preference, 127, 105442.https://doi.org/10.1016/j.foodqual.2025.105442

Manuscript in preparation

- **Ariakpomu, N. C.**, Gill, V., Holmes, M. J., & Ho, P. (not yet submitted). Monitoring Assessor performance with the many-facet Rasch model.

Conference Presentations

- **Ariakpomu, N. C.**, Holmes, M. J., & Ho, P. "Measuring overall difference from a combination of attribute ratings with the many-facet Rasch model" (September 2024). ***11th European Conference on Sensory and Consumer Research-EUROSENSE 2024: A Sense of Global Culture***, Dublin, Ireland. (Poster presentation).

- **Ariakpomu, N. C.**, Gill, V., Holmes, M. J., & Ho, P. "Monitoring assessor performance using a many-facet Rasch approach: a comparison of trained and untrained panels" (August 2025). ***16th Pangborn Sensory Science Symposium: Connecting Sense and Minds***, Philadelphia, USA. (Poster presentation).

Conference Attendance (No Presentation)

- **IFT FIRST Annual Conference**: focused on the latest global advancements in food technology, innovation, and consumer trends (July 2024), Chicago, USA.
- **IFST Oxford Food Summit 2025 – *Designing adaptive food systems for sustainable nutrition***: focused on exploring innovative strategies to design sustainable and adaptive food systems (July 2025), Oxford, UK.
- **Growth Asia Summit:** focused on food, beverage and nutrition trends and innovation in the Asian market (July 2025), Singapore.

- **International Society of Neurogastronomy (ISN) 2025 Symposium**: focused on the interdisciplinary science of flavour perception, integrating neuroscience and culinary arts, to explore how taste and smell influence food preferences and well-being (August 2025), Philadelphia, USA.

Professional Development

- **Your Entrepreneurs Scheme (YES23) Competition** (2023): A bootcamp on commercialising research, in partnership with The University of Nottingham's Haydn Green Institute for Innovation and Entrepreneurship.
- **100 Black Women Professors Now** (2024): Career accelerator program offered by the University of Leeds and Women's Higher Education Network (WHEN).
- **Industry Engagement:** site visits to food manufacturing companies (such as Cranswick Plc and Premier Foods), and discussions with consultants from Cambridge Market Research, Sense:lab, Campden BRI, MMR Sensory, Compusense, and independent consultants. These interactions offered practical insights that were instrumental in establishing the study's relevance and applicability to industry practice.

Dedication

In loving memory of my beloved dad, Dr. Ray Unamma who taught me to always fully experience education rather than just cruise by.

I let the PhD pass through me as well.

Acknowledgements

I am perpetually grateful to God Almighty for His abundant grace and favour that daily saturate my life. My PhD journey has been a tremendous opportunity for growth, supported by many incredible individuals and organisations.

I am especially thankful to the Commonwealth Scholarships Commission in the UK for recognising my leadership potential and funding my doctoral studies. It has made a profound impact on my life; one I intend to carry forward. I am also grateful to Michael Okpara University of Agriculture, Umudike (MOUAU), for nominating me for the scholarship and for believing in my potential as a change-maker.

To the inspirational women I have met through the 100 Black Women Professors Now programme, thank you for your empowering words and the supportive community you have created. To OD&PL (Organisational Development and Professional Learning), Skills@Library, and IT Training, thank you for the invaluable knowledge and skills I gained through your many programmes at the University of Leeds.

To my supervisors, Dr. Peter Ho and Dr. Melvin Holmes, I am deeply grateful for your patient support, invaluable guidance, and thoughtful feedback throughout my doctoral studies, especially for grounding me in the statistical rigour that underpins my research. Your insight and encouragement challenged me to think more critically, and I am better for it. I am also very thankful to Ian and Miles for providing technical support during the sensory evaluation studies, and to the study participants for always showing up and lending us your time and attention.

To Lauren Rogers, Simon Woods, Bryson Bolton, and Dr. Stella Salisu, thank you for your mentorship and for giving me a window into the world of food industry practices, which helped shape the relevance of my research and career plans.

A very special note of appreciation goes to the Food Admin team. Words are not enough to express how thankful I am for your incredible support throughout my time at the University of Leeds. Jenna, Gita, George, Sarah, Katelyn, Matthew and Catherine, thank you so much for always listening and for your responsive, generous help. To my office mates at Room 1.07; Sadia, Dolapo, Arig and Gizem, thank you for making this journey such a memorable experience. To Teresa,

Blessing, Ann, Chinwe, Flora, Nadia, and Kaya, thank you for your beautiful friendship and for being such an important support system.

My beloved family have always been my greatest cheerleaders. From the voice of my late dad, always reminding me, “Nnenna, you have to let the school pass through you too”, to the pride I see in my siblings, even when they make fun of their baby sis, thank you! Thank you, Emeka, for taking on the responsibility of putting me through school, among everything else. Thank you, Mma, Eddie, Aunty Chinenye, and Uncle Tunde, for always being there for me. Thank you, Mummylistic! I know your prayers are working in my life, and I hope I continue to make you proud.

Finally, to my dearest husband Clifford, my very own personal industry standards and pragmatism consultant, I am endlessly grateful. The support from you and our boys has been my greatest source of inspiration and motivation. We have conquered so many challenges and beaten so many odds together, and I am honoured to have you by my side. Thank you so much. We have done it again!

To everyone whose names I could not include here, please know that your kindness, support, and encouragement are deeply appreciated and will always be cherished.

This document contains internal hyperlinks (e.g., in the table of contents and cross-references to figures, tables, sections, and footnotes). For the best navigation and reading experience, it is recommended to view the PDF in **Adobe Acrobat Reader**, as it currently supports returning to the previous location after clicking a hyperlink.

To return, press **Alt + Left Arrow (←)** on Windows or **Command + Left Arrow (←)** on Mac.

Abstract

Traditional discrimination methods either provide holistic product difference scores or focus on specific sensory attributes, often requiring multiple tests to capture both qualitative and quantitative insights. While aggregate-based analyses like ANOVA can statistically adjust product comparisons for assessor effects, they do not identify which individual assessors exhibit problematic rating behaviours, such as using limited parts of the scales or being too lenient or severe. Obtaining these diagnostic insights to guide targeted interventions (e.g., retraining or panel refinement) requires separate analyses that are not integrated into the standard discrimination testing framework.

This research explores the application of a Many-Facet Rasch Model (MFRM) as a diagnostic and analytical tool in sensory difference testing. MFRM addresses these challenges by estimating a single latent measure of overall product difference from combined ratings of multiple attributes, while simultaneously adjusting for individual differences in scale use. It also offers integrated quality control metrics that support panel diagnostics and highlight the discriminative value of individual attributes.

Across three studies, trained and untrained panels evaluated the intensity of various sensory attributes in three different food products. Rasch-derived overall difference measures aligned closely with results from the Difference-from-Control (DFC) overall difference test. Wright maps visualised the relative difficulty of perceiving attributes and the rating tendencies of individual assessors, while fit statistics and residual analyses revealed the contributions of individual attributes to perceived product differences and systematic rating patterns. MFRM further identified distinct types of individual scale-use bias, supporting targeted assessor training.

This study establishes the MFRM as a scalable, more insightful approach for sensory data analysis, with applications in quality control, product development, and panel management. Further research is encouraged to explore its utility across broader sensory and consumer testing contexts.

Keywords: Many-Facet Rasch Model (MFRM), Sensory difference, Attribute discrimination, Difference-from-Control (DFC), Assessor performance monitoring, Scale-use bias, Sensory data analysis, Quality control, Product development, ANOVA.

Table of Contents

List of Tables

List of Figures

Ethics and Data Statement

The data for this research were collected from three sensory evaluation studies. Ethical approval for the involvement of human subjects was granted by the MaPS and Engineering Joint Faculty Research Ethics Committee, and the Business, Environment and Social Sciences Faculty Research Ethics Committee at the University of Leeds. The corresponding datasets are available in the Research Data Leeds Repository, with dataset and ethics approval references provided below.

Chapter	Dataset	Ethics Approval Reference
Chapter 4	Sensory attribute and difference from control ratings of Jaffa cakes (Ariakpomu et al., 2024). https://doi.org/10.5518/1484	AREA FREC 2023-0433-496
Chapter 5	Sensory attribute ratings of chocolate spreads. (Gill et al., 2024). https://doi.org/10.5518/1483	MEEC 15 -003
Chapter 6	Sensory attribute and difference from control ratings of tomato soup. (Ariakpomu et al., 2025a). https://doi.org/10.5518/1658	BESS+FREC - 2024 0433-2568

Copies of the ethics approval letters are provided in **Appendix A**

List of Abbreviations

ANOVA	Analysis of Variance
AR test	Attribute Rating test
CATA	Check-All-That-Apply
CLTs	Central Location Tests
DFC	Difference-from-Control
DFCM	Difference-from-Control (Rasch) Model
DIF/DFF	Differential Item Functioning / Differential Facet Functioning
HPLC	High Performance Liquid Chromatography
HUTs	Home Use Tests
INFIT	Inlier-sensitive Fit
ISO	International Organisation for Standardisation
JMLE	Joint Maximum Likelihood Estimation
LID	Local Item Dependence
LMS	Labelled Magnitude Scale
GC-MS	Gas Chromatography-Mass Spectrometry
gLMS	Generalised Labelled Magnitude Scale
MANOVA	Multivariate Analysis of Variance
MFRM	Many-Facet Rasch Model
NA	Not Applicable
OUTFIT	Outlier-sensitive fit
OUTFIT Mnsq	Outlier-sensitive Mean Square
PCA	Principal Component Analysis
PCAR	Principal Component Analysis of Residuals
pg.	Page number (in-text cross references)
PT Measure	Point-Biserial Correlation Measure
QA/QC	Quality Assurance / Quality Control
QDA	Quantitative Descriptive Analysis
RATA	Rate-All-That-Apply
RCBD	Randomised Complete Block Design
S.D	Standard Deviation
S.E	Standard Error
SR/ROR	Single Rater - Rest of Rater correlation
TCATA	Temporal-Check-All-That-Apply
TIM	Total Intensity Measure
Tukey’s HSD	Tukey’s Honestly Significant Difference
VR / AR	Virtual Reality / Augmented Reality
+VE	Positive
-VE	Negative

Chapter 1

Introduction

1.1 Background of study

The sensory evaluation of food products plays a critical role in the food industry, informing product development, quality control, marketing, regulatory compliance, and consumer satisfaction (Stone et al., 2012; Heymann, 2019; Moskowitz & Meiselman, 2020). Accurate and reliable product characterisation, through methods such as descriptive analysis or sensory profiling, and sensory difference testing, help manufacturers understand how product variations influence consumer perception and acceptance.

Current difference testing methods involve trade-offs between qualitative and quantitative insights. For example, triangle tests reveal if products differ overall, while attribute-specific tests like paired comparison (2-AFC) tests identify which product differs with respect to a single attribute (e.g., sweetness). These approaches provide qualitative information about the existence of differences but not their magnitude. In contrast, the Difference-from-Control (DFC) test quantifies the overall magnitude of difference between products but does not indicate which attributes are responsible. As a result, multiple tests and statistical analyses are often required to gather both qualitative and quantitative insights, (Rogers, 2017; Higgins & Hayes, 2020), making the process time-consuming and resource-intensive.

Attribute Rating (AR) tests on the other hand, typically part of descriptive analysis enable the simultaneous rating of multiple attributes. However, interpreting overall product differences from AR data requires complex multivariate analysis. Moreover, these methods rely on extensively trained panels, which are costly to maintain since they must remain motivated and consistently calibrated to rating scales over time (Raithatha & Rogers, 2018; Moskowitz & Meiselman, 2020; Meilgaard et al., 2025).

Analysing sensory data presents further challenges related to reliability and validity (Kemp et al., 2018). Because panel ratings are typically aggregated, it is difficult to

estimate the true differences between samples without confounding effects from individual assessor variability, an issue that even rigorous training cannot fully eliminate (Næs, 1990; Romano et al., 2008). Statistical models such as Analysis of Variance (ANOVA) are used to account for these effects, but they rely on assumptions that may not always hold, and they can still be influenced by inconsistent rating styles.

Monitoring assessor performance is essential for identifying individuals whose responses deviate from panel expectations and determining when additional training or removal from the panel is necessary. Individual-level analyses, such as assessor-specific ANOVAs, are often used to evaluate panel performance. However, these approaches rely on aggregated data (e.g., mean scores across replicates), which can obscure subtle inconsistencies in rating behaviour or individual variability. In practice, effective monitoring often depends heavily on the expertise of the panel leader in recognising these deviations and implementing corrective actions. Detecting inconsistencies can require multiple layers of analysis and visualisation, which may be time-consuming and slow decision-making in commercial environments (Raithatha & Rogers, 2018), despite the availability of sensory analysis software to automate parts of the process (Fuentes et al., 2021; Sipos et al., 2021).

Beyond monitoring, addressing individual differences in sensory responses during data analysis presents additional challenges. Although some studies have proposed methods to account for these differences, such as adjusting for overall scale use through the assessor model (Romano et al., 2008) or evaluating consumer inconsistency using Kendall's rank correlation coefficient between paired scales (Sipos et al., 2025) these approaches still involve multiple analytical steps.

Collectively, these challenges underscore a need for analytical techniques that can efficiently integrate product differentiation with attribute diagnostics and panel performance monitoring within a single analytical framework. Addressing this need could streamline sensory workflows, reduce analytical costs, and provide more actionable insights for product developers, quality managers, and panel leaders.

This study explores the application of Rasch analysis, a psychometric modelling approach that analyses data based on individual response patterns (Bond et al.,

2020), in sensory difference testing. Rasch analysis offers the potential to unify intensity ratings across multiple sensory attributes into a single latent measure of overall difference, while simultaneously providing diagnostic insights into attribute contributions and assessor reliability within a unified framework. In this study, intensity ratings were collected for multiple sensory attributes across a range of product samples. These attribute ratings were then combined into an overall difference measure and compared with holistic Difference-from-Control (DFC) scores. Panel performance was assessed concurrently through the Rasch model.

1.2 Research hypothesis

Hypothesis 1:

Rasch analysis of intensity ratings for multiple sensory attributes can provide a comprehensive estimate of the overall difference between food product samples, based on the ratings provided by the panel, and enable identification of the attributes that contribute most to these differences.

Rationale:

Rasch models combine multiple observable items (e.g., test questions) to estimate unobservable latent variables (e.g., mathematical ability). When applied to sensory evaluation, the model can combine intensity ratings of multiple sensory attributes (e.g., sweetness or sponginess) to derive a single latent measure representing the overall difference between products. This approach enables (i) the detection of whether a significant difference exists between samples, (ii) quantifies the magnitude of difference if one exists, and (iii) determines the relative contribution of individual attributes to the perceived differences, all based on the panel ratings from a single sensory test. In contrast to conventional approaches, which often rely on multiple separate tests and analyses, Rasch analysis provides a more streamlined, cost effective, and diagnostic tool for sensory analysts.

Hypothesis 2

Using Rasch analysis to monitor assessor performance enables earlier identification of assessors needing additional training, thereby reducing overall training time and resources.

Rationale:

Rasch models inherently account for individual differences in rating tendencies (e.g., severity or leniency) and include diagnostic tools such as residual fit statistics (e.g. outfit mean square) and Wright maps that visualise the overall structure of the data. These features enable individualised evaluation of assessor performance and consistency relative to the panel expectations. This approach allows for a rapid overview of individual assessor performance and offers insights to support early and targeted training interventions, all within the same integrated analysis used to evaluate overall product differences (as described in Hypothesis 1). As a result, it has the potential to reduce both training time and resource demands. In comparison, conventional approaches do not adjust for individual rating behaviour. Instead, they emphasise rigorous training to standardise assessors as objective rating instruments and depend on multiple separate analyses to evaluate performance.

1.3 Research aims

This study explores the potential of Rasch analysis to provide a streamlined, integrated diagnostic framework for sensory difference testing, enabling simultaneous evaluation of overall product differences and individual assessor reliability.

The specific objectives are:

1. To demonstrate the use of Rasch analysis in estimating an overall difference (latent variable) between food product samples from a combination of sensory attribute intensity ratings.
    - Collect sensory data using the Difference-from-Control (DFC) test method.
    - Collect attribute intensity ratings for multiple sensory attributes on the same food samples and using the same group of assessors.
    - Compare the DFC-derived overall difference results with Rasch-generated measures of the *Overall Difference* as a latent variable.
2. To demonstrate how Rasch model quality control features can be used to assess the reliability of assessors as objective measurement instruments.

- Evaluate individual assessor performance using Rasch-based diagnostics, such as fit statistics and Wright maps.
- Identify assessors whose ratings deviate from panel expectations to inform potential retraining.

3. To assess the reproducibility of the proposed method in a new context using an integrated approach.
    - A validation study conducted to assess both assessor performance and overall product differences within a single Rasch analysis, applied in a context that closely reflects practical conditions found in food production settings.

1.4 Thesis structure

This thesis consists of seven chapters, and a brief overview of the chapters following **Chapter 1** (this introductory chapter) is provided below.

Chapter 2 reviews the literature on sensory difference testing methods, highlighting the role of humans as measuring instruments and the challenges posed by individual variability. It then introduces Rasch models, illustrating their current applications and their potential relevance to sensory quality control.

Chapter 3 outlines the general methodologies used in this study for Rasch and statistical analysis. It explains how Rasch measurement approaches are applied across three research themes, which will be discussed in Chapters 4 to 6.

Chapters 4 to 6 outline the specific sensory methodologies used in the three sub-studies, and present the results and discussions for each, exploring the different applications of Rasch analysis in sensory difference testing and quality control. Specifically:

Chapter 4 focuses on using the Many-Facet Rasch Model (MFRM) to measure the overall difference between samples through a holistic Rasch measure, termed the Total Intensity Measure (TIM). TIM is estimated for each sample based on a combination of intensity ratings from five sensory attributes. Sensory attribute ratings and Difference from Control (DFC) ratings of Jaffa cakes were used for the

study, and the results from TIM were compared to those from DFC. Additionally, the chapter discusses how Rasch quality control statistics provide deeper insights into the contribution of each attribute to the overall difference latent variable, as well as how easy or challenging it was for the panel to evaluate an attribute, highlighting the added benefits of the TIM method over traditional overall difference methods.

Chapter 5 focuses on monitoring assessor and panel performance using the MFRM approach. It compares trained and untrained panels based on their sensory attribute ratings on chocolate spreads. Results from Rasch quality control statistics were compared with those from ANOVA-based methods, alongside response distribution plots of individual ratings for each attribute. Based on the insights from the analysis, a subset of better-performing untrained assessors was identified and compared with the trained panel.

Chapter 6 demonstrates how a Rasch approach can streamline sensory quality programmes. In summary, the Total Intensity Measure (TIM) was used to assess the overall difference between tomato soup samples based on a combination of eighteen sensory attributes, while also providing insights into each attribute's contribution to the overall difference latent variable, and which of the attributes were easy or challenging to evaluate. Within the same analysis, Rasch quality control statistics were also used to monitor assessor performance and identify areas for targeted training. A subset of the most consistent assessors, as identified by the model, was then selected to run the study, demonstrating how the Rasch approach can aid assessor selection and guide targeted training. The TIM overall difference results were compared with those from the Difference from Control (DFC) test on the same samples to validate the findings while addressing limitations identified in the previous chapters.

Chapter 7 summarises the key findings of the thesis and their implications, concluding with recommendations for areas where future research could build on the research findings.

Chapter 2

Literature Review

2.1 Overview of Sensory evaluation

The field of sensory science and consumer research has evolved over the decades through interrelated generations of scientific work. Foundational contributions from academic figures such as Rose Marie Pangborn and Maynard Amerine, along with the growth of the food industry after World War II, particularly in sectors like wine and brewing, have shaped its trajectory. Shapin (2016), Lahne and Spackman (2018) and Moskowitz and Meiselman (2020) provide detailed historical accounts of this evolution, complemented by a more personal narrative in Heymann (2019) and an account of developments and ongoing challenges in the field by Meiselman et al. (2022).

Sensory evaluation, as originally defined by the Institute of Food Technologists (IFT), Chicago in 1975 (Heymann, 2019), refers to the scientific methods used to evoke, measure, analyse and interpret human responses to the properties of foods as perceived by the human senses including taste, smell, touch, sight, and hearing. This definition was later expanded to include the role of the trigeminal nerves, which contribute to sensations such as heat, cooling, and irritation (Lawless & Heymann, 2010; Stone et al., 2012).

In recent times the field has become increasingly interdisciplinary and is now applied across a broad range of consumer products beyond the food industry (Kemp et al., 2018; Heymann, 2019; Meiselman et al., 2022; Jaeger et al., 2025; Meilgaard et al., 2025). These include, pharmaceuticals (Mohamed-Ahmed et al., 2016; Guedes et al., 2021; Clapham et al., 2023), personal and household care products (Sanderson & Hollowood, 2017; Deubler et al., 2022; Turek & Kowalska, 2024), automobiles (Poirson et al., 2010; Verriele et al., 2012; Othman et al., 2021; Fuchs et al., 2022), fashion & textiles (Ghalachyan et al., 2024; Üren, 2024), and even pet foods evaluated using animal assessors (Li et al., 2018; Lema Almeida et al., 2022; Rogues et al., 2022; Calderón et al., 2024; Le Guillas et al., 2024).

Sensory evaluation of foods supports a wide range of industrial applications, including product optimisation, shelf-life and stability testing, quality control, market audits, benchmarking, and substantiating legal or advertising claims (Stone et al., 2012). It aims to derive objective insights from inherently subjective human perceptions, providing essential guidance for commercial decision-making. Guided by seminal foundational texts, cited here in their latest editions (Amerine et al., 1965; Muñoz et al., 1992; Lawless & Heymann, 2010; Næs et al., 2010; Stone et al., 2020; Meilgaard et al., 2025), sensory evaluation methods have been developed and refined to reflect current best practices. These methods apply principles of experimental design and statistical analysis, enabling sensory professionals to make valid inferences and generate actionable insights about food products. Sensory test methods are broadly classified as objective or subjective. **Objective** methods aim to characterise the sensory attributes of products and typically rely on trained or expert panels; these include discrimination and descriptive tests. **Subjective** methods assess how product changes affect consumer perception and generally involve larger panels of untrained assessors or consumers, such as in preference and acceptance tests. Marques et al. (2022) provides a comprehensive review of both classical and emerging sensory evaluation methodologies within the food and beverage industry.

2.2 Sensory evaluation in quality control

According to Meilgaard et al. (2025), "*sensory quality*" refers to the procedures implemented to ensure that products leaving a manufacturing facility meet established design parameters and consumer expectations regarding sensory attributes. It encompasses both proactive sensory quality assurance, aimed at preventing defects, and reactive sensory quality control, which focuses on identifying and correcting them. In sensory quality programmes, the product-oriented methods i.e. discrimination and descriptive tests are typically employed. Several researchers (Muñoz et al., 1992; Costell, 2002; Muñoz, 2002; Rogers, 2017; Meiselman et al., 2022; Meilgaard et al., 2025) have identified discrimination tests such as the "In Out" and Difference from Control (DFC) methods, along with attribute descriptive tests, as effective approaches in this context.

2.2.1 Discrimination Tests

Discrimination tests are one of the key product-oriented sensory methods used in quality control programmes, as mentioned earlier. They aim to evaluate whether perceptible differences exists between two or more products and have been classified into overall and attribute-specific tests (Lawless & Heymann, 2010; Bi, 2015; Rogers, 2017; Meilgaard et al., 2025), depending on whether the test specifies the nature of the difference in advance. These methods are also referred to as "Unspecified and Specified" difference methods (Amerine et al., 1965; Bi, 2015) or "Non-directional and Directional" discrimination methods (Lawless & Heymann, 2010) respectively.

Overall difference tests require assessors to identify if a sample among a set differs from the others, without specifying the attribute of interest. These include:

- **Triangle test**: Identify the odd sample out from three samples (two identical, one different).
- **Tetrad test**: Evaluate the four samples and group them into two groups of two based on similarity.
- **Duo-Trio test**: Identify which of two coded samples matches a known reference.
- **Two-Out-of-Five test**: Out of the five samples presented, three are of one kind and two are of another. Identify the two samples that are different from the other three.
- **Same–Different/simple difference test**: Judge whether two samples are the same or different.
- **Difference-from-control (DFC)**: Rate how much a test sample differs from a control or reference sample on a specified rating scale.

Attribute-specific tests focus the attention of assessors on a particular sensory characteristic, ignoring other differences. These include as examples:

- **Paired comparison**, 2-AFC: which of the samples is sweeter?
- **Alternative forced choice methods** (3-AFC, 4-AFC): which of the samples is the sweetest?

- **A-NOT A**: in terms of saltiness, is this sample the same as A or Not A? The In-Out test is a variation of the A-NOT A method, in which assessors classify samples as either within or outside the acceptable range of variation defined by a target product (Muñoz, 2002; Meilgaard et al., 2025).
- **Rank test**: Rank this samples in order of increasing intensity of an attribute (sweetness, saltiness, etc).
- **Attribute rating test**: Rate the samples on a specified scale according to the intensity of an attribute (sweetness, saltiness, etc).

Overall difference tests rely on a holistic comparison strategy, require relatively minimal training, and do not indicate which specific sensory attributes differ. Attribute-specific tests, on the other hand, focus on a single attribute but demand greater cognitive effort, adequate sensitivity to that attribute, and often additional assessor training.

Detailed test procedures for each test are provided in several texts including (Lawless & Heymann, 2010; Stone et al., 2012; Rogers, 2017; Meilgaard et al., 2025). Depending on the objective, the goal of most discrimination tests (except for ranking and rating tests) may be to demonstrate that products differ (difference testing), or to establish that they are similar enough to be used interchangeably (similarity testing). In similarity testing, the same test designs are used, but the statistical hypotheses are reversed to determine whether any sensory differences are small enough to be regarded as negligible. In some cases, such as with the DFC test, the goal extends to quantifying the magnitude of difference, providing more actionable insights beyond simple binary responses.

The DFC test, originally introduced by (Aust et al., 1985) as the Degree of Difference (DoD) test, is valued for its simplicity and unique ability to capture not only the presence of a perceptible difference but also the magnitude of that difference relative to a control sample. This sets it apart from tests like the Triangle and Duo-Trio, which only produce binary outcomes, and is particularly useful for tracking batch-to-batch variation in heterogeneous products.

However, while these methods can identify whether a perceptible difference exists, they do not provide insight into why products differ in terms of attributes. Attribute-

specific tests, for instance, indicate the presence of a difference in only a single attribute but neither quantify the magnitude of that difference nor capture multiple contributing factors, unless separate tests are conducted per attribute. In contrast, the DFC test offers a quantitative measure of the magnitude of perceived differences but still does not reveal the specific sensory attributes responsible for those differences.

To address this limitation, several studies have explored combining the DFC test with additional qualitative methods. Rogers (2017) suggests including a comment section to capture assessor perceptions of what might be causing the differences. Compusense (2020), a white paper on quality control using the DFC test, suggested incorporating check-all-that-apply (CATA) follow-up questionnaires to improve manufacturers' chances of identifying product faults. Similarly, Higgins and Hayes (2020) combined CATA questions with an open-ended comment box to further characterise differences in beer samples.

Despite these enhancements, the resulting attribute insights remain qualitative, providing only basic information about the presence or absence of certain attributes. Although statistical tests such as the Cochran's Q can be applied to the CATA responses to determine which attributes are selected significantly more frequently across samples (Meyners et al., 2013; Meyners & Hasted, 2021), thereby identifying attributes that likely differ between products, this approach remains fundamentally frequency-based. It does not directly measure the intensity or the relative contribution of individual attributes to the overall perceived difference, highlighting a gap that warrants further methodological development.

2.2.2 Attribute Rating (AR) Tests / Descriptive analysis

According to Muñoz et al. (1992), Attribute Rating (AR) tests, are one of the most powerful tools for assessing the sensory quality of products. They are central to descriptive analysis or sensory profiling methods, enabling the quantification of specific sensory characteristics using intensity rating scales. Over time, the development of descriptive analysis has reflected a continuous effort to overcome limitations in panel reliability, objectivity, and comparability across time and products. The major methods in this category illustrate this progression:

- The **Flavour profile™** method (Caul, 1957) was one of the earliest techniques. It used a consensus-based approach, where assessors discussed and agreed on the description and intensity of aroma, flavour, and aftertaste. However, its reliance on panel consensus limited reproducibility and masked individual variability in perception, which posed challenges for statistical analysis and scalability.
- In response, **Quantitative Descriptive Analysis (QDA™)** was developed in 1974 (Stone et al., 2004). This method shifted from group consensus on attribute intensities to independent evaluation, training each assessor to rate the consensus-generated sensory attributes individually on unstructured line scales. This improved statistical robustness and allowed for more objective data collection. However, the method lacked standardised reference points, leading to potential inconsistencies between panels and over time.
- As a refinement to both the Flavour profile™ and the QDA™, the **Spectrum™ method** developed by Gail Vance Civille and officially named "Spectrum Descriptive Analysis" in 1986 (Civille & Osdoba, 2020; Meilgaard et al., 2025), introduced anchored rating scales based on physical and conceptual reference standards. This ensured improved calibration and consistency, making it especially suitable for long-term product tracking and cross-laboratory comparisons. Spectrum retains the independent evaluation of QDA but adds rigor through standardised training and reference materials.

This evolution reflects a deliberate shift toward methods that balance individual sensitivity, panel consistency, and data reproducibility, each stage refining the scientific reliability of sensory measurement for quality control and product development. Among these, Quantitative Descriptive Analysis (QDA) and the Spectrum method are widely used in industrial practice (Meiselman et al., 2022) for their structured approaches, which promote consistency, comparability, and reliable interpretation of sensory data.

In sensory quality programs, these methods focus on identifying critical sensory attributes, those known to introduce variability within a product (Muñoz et al., 1992; Meilgaard et al., 2025). Consumer acceptance data are used to establish sensory

specifications and define acceptable limits for these attributes, ensuring the product consistently meets consumer expectations.

Highly trained or expert panels are essential for this approach, which makes the method resource intensive. Assessors must generate a sensory lexicon (a standardised vocabulary of relevant attributes), undergo extensive training and calibration to ensure that the rating scales are used uniformly, and maintain consistency through ongoing performance checks (Raithatha & Rogers, 2018; Meilgaard et al., 2025). As such, while descriptive profiling provides rich quantitative data for decision-making, it can be time-consuming and costly due to its complexity and reliance on skilled personnel (Næs, 1990; Ares, 2015; Raithatha & Rogers, 2018; Moskowitz & Meiselman, 2020; Torrico et al., 2023; Meilgaard et al., 2025).

To accelerate product development and deliver innovations that meet consumer expectations with minimal training, rapid sensory profiling methods emerged in the 2000s.

- **Flash Profiling (FP)** (Delarue & Sieffermann, 2004) allows assessors to generate their own descriptive terms and rank products based on the perceived intensity of these attributes.
- **Napping®** (Pagès, 2005) and **Sorted Napping** (Pagès et al., 2010) , are projective mapping techniques in which assessors place products on a 2D surface based on perceived similarities; the latter adds a grouping step for similar items.
- **Polarised Sensory Positioning (PSP) (**Teillet et al., 2010**)** where assessors compare test products to a small set of selected reference products (poles), and rate how similar or different each product is to each pole. This results in a holistic, perceptual map that visualises the overall sensory relationships among all products and poles.
- **Check-All-That-Apply (CATA)** (Ares et al., 2010), presents assessors with a predefined list of sensory attributes, from which they check all the attributes perceived in the product being evaluated.

Time-dependent methods were also developed to capture dynamic sensory perception.

- **Temporal Dominance of Sensations (TDS)** (Pineau et al., 2009) requires assessors to continuously select the most dominant sensory attribute over time.
- **Temporal Check-All-That-Apply (TCATA)** (Castura et al., 2016) an evolution of the CATA but similar to the TDS requires assessors to check all attributes that apply to the product at different points in time instead of the most dominant attribute.

Of all the rapid methods, CATA has become the most widely applied (Jaeger et al., 2015; Vidal et al., 2018; Ruiz-Capillas & Herrero, 2021; Kim et al., 2023) due to its simplicity and consumer-centric approach. However, because it does not capture intensity data, extensions like the Rate-All-That-Apply (RATA) (Ares et al., 2014) were developed.

- **RATA** enables 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, offering a practical yet quantitative alternative to traditional profiling. It has been shown to improve product discrimination compared to CATA (Ares et al., 2014; Reinbach et al., 2014).

Despite the growing interest in rapid methods, attribute rating tests have remained essential in sensory quality control. Another of their key advantage over the Difference-from-Control (DFC) test is that samples are assessed independently, without requiring comparison to a reference. This makes them less cognitively demanding for assessors and reduces fatigue, particularly when a large number of samples must be evaluated. In contrast, DFC tests, although useful for quantifying overall product difference, can be resource-intensive, especially when multiple products are involved. Even when test sessions are spread out over time, they often require a greater time commitment from both assessors and researchers.

Recent methods like RATA have increased the accessibility and consumer relevance of sensory profiling and attribute rating tests, but trade-offs remain. Attribute rating provides detailed quantitative data on individual attributes but does not directly measure overall product difference. Instead, overall differences must be inferred through multivariate analysis, which can add complexity to result interpretation. In contrast, DFC tests capture overall product differences but

require additional tests to identify the key sensory drivers. This highlights the ongoing need for an integrated approach that reliably and efficiently captures both overall product differences and the contributions of individual attributes in a reliable, interpretable, and efficient manner.

2.3 Measuring responses in sensory evaluation

At the core of all sensory testing methods is an unavoidable challenge, which is the inherent variability of the human sensory system. Regardless of the evaluation method used, the data collected must remain reliable and interpretable, even though humans serve as the measuring instruments. As the link between technical product development and market realities, the sensory analyst plays a critical role in ensuring that product changes are accurately measured and that findings are translated into insights that reflect real consumer experiences and expectations.

2.3.1 Humans as measuring instruments

A defining feature of all sensory evaluation methods is the use of humans, commonly referred to as subjects, assessors, panellists, judges, raters, participants, or tasters, or a group of humans known collectively as a "sensory panel" (ISO 8586:2023 - British Standards Institution (2023)), as the measuring instruments. Unlike mechanical or digital devices, human responses are influenced by a range of internal and external variables, making individual variability inherent to sensory science. This variability stems from differences in past experiences, sensory acuity, health status, and contextual factors (Stone et al., 2012; Meilgaard et al., 2025), as well as from genetic differences in taste receptor genes that create fundamentally different sensory experiences across individuals (Bartoshuk et al., 2005; Feeney et al., 2011). These sources of variability contribute to inconsistencies and background noise in sensory data, making the assessment and management of panel performance a key focus in sensory evaluation (Sipos et al., 2021).

Even with rigorous training, assessors vary not only from each other but also within themselves over time Næs et al. (2010); Stone et al. (2012); and Sipos et al. (2021). This inherent variability highlights the importance of well-designed testing protocols, assessor calibration, and appropriate statistical tools to reduce noise and support valid interpretations. Such fluctuations in perception beyond differences in the

products themselves are an accepted and inherent part of sensory evaluation (Meilgaard et al., 2025).

2.3.2 Individual differences in sensory evaluation

The factors contributing to individual variation and response bias in sensory evaluation, summarised in **Table 2.1**, are broadly classified as intrinsic (inherent to the individual) and extrinsic (external factors contributing to response bias). Intrinsic factors influence how individuals perceive and interpret sensory stimuli and often persist despite training. Extrinsic factors, by contrast, stem from the testing environment or methodology, such as sample presentation, location, and questionnaire design, and are generally easier to control through clear protocols.

Both types of factors have significant implications for the design of sensory tests and the interpretation of results. Minimising their impact is essential for obtaining valid and reliable data. Current approaches to address these include assessor training and calibration to the rating scale, addressing common sources of bias in sample handling, environmental conditions, and test instructions, and using well-structured experimental designs with randomised and balanced presentation (Lawless & Heymann, 2010; Kemp et al., 2018; Meilgaard et al., 2025).

In the wake of increasing globalisation and the post-COVID-19 era, certain factors influencing sensory response have gained prominence, particularly cultural considerations and the influence of contextual testing environments.

Cultural differences can significantly affect how rating scales are used and how assessors interpret the meaning, importance, and intensity of sensory attributes (Lee & Lopetcharat, 2017; Yang & Lee, 2019; Dupas de Matos et al., 2025). As a result, cultural sensitivity has become crucial in the design of questionnaires and rating scales. A global approach to sensory research is encouraged, with growing emphasis on ethical relevance and cultural adaptability to ensure that methods and conclusions remain valid across diverse populations (Muñoz, 2002; Meiselman et al., 2022; Hort, 2024).

Table 2.1. Common sources of individual variation and response bias in sensory evaluation

INTRINSIC FACTORS (Inherent to individuals)	
Factors	**Relevance**
Age	As individuals age, there is a natural decline in the number and function of taste buds and olfactory receptors, leading to diminished taste and smell sensitivity; while children often have heightened sensory sensitivity compared to adults (Guinard, 2000; Issanchou, 2015)
Cognitive ability	Cognitive traits like attention span, learning ability, and memory capacity can influence sensory judgements especially in tests that require mental recalls of stimuli like the Two-Out-of-Five Test (Meilgaard et al., 2025).
Experience	An individual's background including culture, environment, knowledge, and skills affects sensory perception. For example, cultural differences influence culinary experiences and expectations of how food should taste (Ares, 2018).
Genetic predisposition	Genetic differences influence sensory perception (Feeney et al., 2021), as seen in classifications such as supertasters, medium tasters, or non-tasters of phenylthiocarbamide (PTC) and 6-n-propylthiouracil (PROP) (Bartoshuk, 1979).
Health variations	Health conditions can impair sensory perception, such as anosmia (complete loss of smell), parosmia (distorted sense of smell), ageusia (complete loss of taste), or dysgeusia (altered taste) (Parker et al., 2022; The John Hopkins University Hospital, 2023)
Sensory acuity	Individuals vary in their sensitivity to different stimuli. Threshold detection tests are commonly used to screen sensory assessors for acuity. The lowest concentration of a stimulus detectable is called the absolute threshold, while the highest concentration perceivable is the limit of detection (Breslin, 1996; Lawless & Heymann, 2010)
Biological sex	This has been shown to affect sensory capabilities due to hormonal differences and a higher density of taste papillae (Bartoshuk et al., 1994). Research indicates that women generally possess a more acute sense of smell and taste compared to men (Doty & Cameron, 2009).

EXTRINSIC FACTORS (External influences)	
Factors	**Relevance**
Environmental context and test conditions	These include consistency of sample presentation, testing locations, test site setup, ambience of the sensory room and surroundings, carefully designed text questionnaire, etc. British Standards Institution (2019) describes the general guidance for the design of test rooms for sensory evaluation.
Physiological Influences	These include ***adaptation*** or ***mental fatigue*** (a decrease or change in sensitivity to a stimulus due to continued exposure), flavour ***carry-over effects*** (when test samples possess strong, lingering flavours) or ***cross-potentiation*** (where, for instance tasting a sweet sample first heightens the sensitivity to sweetness and results in a higher rating for bitterness in a subsequent bitter sample) (Meilgaard et al., 2025).
Psychological bias	These are the most common sources of bias and refer to systematic mental shortcuts not due to sensory acuity but rather due to external influence on cognitive processes leading to inaccurate ratings and deviations from objective product assessments (Torrico et al., 2023). Common effects (Lawless & Heymann, 2010; Stone et al., 2012; Kemp et al., 2018; Meilgaard et al., 2025) include:
▪ *Dumping effect*	Where assessors assign intensity ratings for perceived but unlisted attributes to a listed attribute instead, effectively inflating that attribute's rating.
▪ *Expectation error*	Prior knowledge about the sample, acquired before or during testing, can influence perception due to preconceived expectations.
▪ *Habituation error*	A tendency to continue giving the same response when the series of test samples presented over time possess gradually increasing or decreasing stimuli.
▪ *Halo effect*	Where the general impression of a product, or the rating for one (dominant) attribute, influences the ratings for other unrelated attributes when multiple attributes are assessed simultaneously.

▪ *Logical error*	When assessors associate several sample characteristics in their minds, like relating green coloured juices with a vegetable or bitter flavour.
▪ *Stimulus error*	Where assessor verdicts are influenced by irrelevant characteristics of the sample itself or its presentation.
▪ *Mutual suggestion*	Where assessor responses are influenced by other assessors' reactions to the product, either through verbal or facial expressions.
▪ *Lack of motivation*	Where assessors are not adequately engaged, interested, or committed to the task of accurately assessing samples.
▪ *Physical condition*	Short-term physical states such as taking medications, smoking, ill health, consuming strongly flavoured food or beverages, or wearing strong perfumes can influence the perception of stimuli.
▪ *Presentation order*	This is a major source of response bias and can manifest in several ways. Presenting a good quality sample just before a poor one may lead to a lower rating for the second sample, and vice versa (***contrast effect***). A good sample presented among poor ones may receive a lower rating than if presented alone (***group effect***). Presenting samples in a particular sequence can lead assessors to anticipate the next sample (***pattern effect***). Assessor attitude may also change over time, with greater anticipation for the first sample and eventual indifference or fatigue toward the last samples (***time-error bias***).
Scale-use bias	This refers to systematic differences in the way assessors use rating scales not reflecting their actual sensory experiences. Differences arise not from the product perception itself but from how individuals choose to express what they perceive. Common forms (Næs, 1990; Myford & Wolfe, 2003; Romano et al., 2008; Kemp et al., 2018; Heymann, 2019) include:
▪ *Level effect*	When assessors consistently rate products higher (***leniency***) or lower (***severity***) on the scale than others.
▪ *Scaling effect*	Where assessors restrict their ratings to a narrow portion of the scale (***restriction of range***), reducing the scale's sensitivity to differences. ***Central tendency*** is a specific case where extreme categories are avoided and responses cluster around the midpoint.
▪ *Extreme response*	Where some assessors use the ends of the scale more than necessary, exaggerating differences.
▪ *Variability effect*	Refers to the internal consistency of assessors when rating repeated evaluations of the same sample.

Environmental context is another evolving area of interest. ***Home Use Tests (HUTs)*** have grown in popularity due to their potential to enhance ecological validity by allowing product evaluations in familiar, real-life settings (Niimi et al., 2022; Torrico et al., 2023). However, HUTs present challenges such as distractions and lack of control over serving conditions, which may compromise data consistency and reliability (Torrico et al., 2020; Giezenaar & Hort, 2021). In contrast, ***Central Location Tests (CLTs)*** typically conducted in sensory booths, provide better control and reduce variability but may introduce response bias, as the artificial setting can influence assessor behaviour (Boutrolle et al., 2007; Hannum et al., 2019).

To bridge these gaps, immersive technologies are being employed to simulate real-life consumption contexts within controlled environments. These include ***Virtual Reality (VR)***, which uses head-mounted displays to present contextual settings (Torrico et al., 2020; Yang et al., 2022); ***Augmented Reality (AR)***, which overlays digital elements such as decorations or a certain ambience within the real-world booth via AR glasses, tablets, or smartphones (Dong et al., 2021); and ***digital immersion***, where the physical testing space is enhanced using 360˚ projection screens, surround sound, and even scent delivery to recreate realistic environments, and include setups like ***immersive walls*** (Hannum et al., 2019) and fully ***immersive rooms*** (Sinesio et al., 2019; Worch et al., 2020; Lichters et al., 2021). These approaches allow for context-relevant testing, without compromising experimental control. Comprehensive reviews of these technologies are available in Fuentes et al. (2021), Giezenaar and Hort (2021), Chai et al. (2022), Torrico et al. (2023)_ENREF_76, and Cosme et al. (2025).

Notably, most research to date has focused on the use of these technologies in affective, consumer-oriented testing, rather than product-focused analytical tests typically used within sensory quality programmes. However, enhancing ecological validity could also support product characterisation, particularly during product development, where understanding product performance in realistic contexts is essential (Ares & Varela, 2017).

Despite advances in sensory methods, critical challenges remain, particularly individual variability and response bias linked to the use of rating scales (Ares, 2018;

Hannum et al., 2019). Tomic et al. (2010) further highlight issues with evaluating panel proficiency i.e., the ability of multiple sensory panels to consistently and accurately evaluate products, across different geographical or global locations. Such inconsistencies can undermine the comparability and harmonisation of results, which are crucial for informed product development. These issues continue to affect the precision and interpretability of sensory data, especially as testing shifts toward more naturalistic settings and diverse consumer populations. Overcoming them requires sophisticated analytical techniques that can distinguish true product differences from variation caused by assessors (Romano et al., 2008) or contextual factors, thereby enhancing the robustness of sensory evaluations and producing results that better reflect real-world consumer perceptions.

2.4 Some considerations in sensory evaluation for quality control

There has been ongoing debate about the use of trained or untrained panels for sensory quality assessment of products, especially with regards to obtaining relevant and representative results that align with consumer expectations (Meiselman, 2013; Ares & Varela, 2017; Moskowitz, 2017). Also, concerns regarding the reliability and validity of results derived from sensory panel assessments continue to be a significant issue within the industry (Raithatha & Rogers, 2018).

2.4.1 The Trained - Untrained panel spectrum

Traditionally, sensory panels have been clearly divided into two categories: trained panels, often treated as analytical instruments, functioning like machines and expected to provide predictable and repeatable data without the influence of personal preference; and consumer panels, valued for their subjective judgements that reflect real world consumer experiences and are usually based on liking and emotional response rather than objective analysis. The two types of panels have typically been kept separate (Meiselman, 2013; Ares & Varela, 2017).

According to ISO 8586:2023 - British Standards Institution (2023), **trained assessors** are screened for sensory acuity relevant to the attributes under evaluation, trained in specific sensory methods (or multiple methods), and maintained over time through follow-up training and validation involving practice with product attributes and rating scales. With continued experience and

demonstrated sensitivity, these individuals become **expert sensory assessors**, capable of delivering consistent, repeatable sensory assessments across various products.

Outside of this classification are **untrained assessors**, who could be **naïve assessors** with no prior sensory evaluation experience, as well as **initiated or experienced assessors** who may have some exposure to sensory testing methods but lack formal training. Somewhere between these groups are **"semi-trained" assessors**, typically industry employees or co-workers who have received some familiarisation with the products and methods but lack the extensive training of a trained panel. This group appears to have evolved as a practical compromise in contexts where time and resources are limited.

Since Meiselman's decade-long prediction that the line between trained and untrained or consumer panels would become increasingly blurred (Meiselman, 2013), this shift appears to be materialising. Ares and Varela (2017) and Moskowitz (2017) make a strong case for both panel types, arguing that the choice between them should depend on the test objective. Both are valuable tools for sensory quality control and product development.

Several studies have shown that untrained assessors can effectively carry out analytical attribute difference tests, particularly when using alternative rapid methods (Giacalone & Hedelund, 2016; Mello et al., 2019; Barton et al., 2020; Maheeka et al., 2021; Wang et al., 2022; Xiangli et al., 2024). Across these studies, a consistent finding is that trained panels tend to use more technical and precise descriptors, whereas untrained or semi-trained assessors rely more on hedonic or general (umbrella) terms.

Interestingly, when comparing performance, trained panels are not always more discriminative overall. Their superior sensitivity emerges primarily for attributes with low detection thresholds which they have been specifically trained to notice. However, this sensitivity is often limited to those attributes. As Chollet et al. (2005) and Ares and Varela (2017) argue, this perceptual advantage doesn't necessarily generalise to stimuli outside their training.

Trained panels are also taught to use more complex rating scales, such as unstructured line scales, often with minimal verbal anchors (discussed in the next section). These scales are assumed to allow for finer discrimination and are treated as yielding interval-level data (Lawless & Heymann, 2010; Meilgaard et al., 2025). In contrast, untrained assessors are expected to produce more variable results. This is partly due to a lack of consensus on where attribute intensities lie along the scale, and partly because their responses are influenced by personal experiences and preferences. Without a shared frame of reference for scaling, different assessors may interpret the same attribute in different ways. As Antmann et al. (2011) showed in their cross-cultural study of creaminess perception, consumers differed considerably in how they understood and rated this attribute, even when evaluating identical samples. Similarly, Ares et al. (2011) reported that consumer panels used unstructured line scales inconsistently, leading to low reliability in texture-intensity ratings of dairy desserts. When results from trained and untrained panels are averaged, however, the rank order of differences between samples is often similar (Worch et al., 2010; Xiangli et al., 2024).

That said, averaging consumer panel intensity scores should be approached with caution. Their inconsistent, heterogeneous use of scale and other scale-related effects can compromise data reliability (Ares & Varela, 2017; Hannum et al., 2019).

Familiarity with the product being evaluated has also been shown to increase sensitivity to specific attributes (Moskowitz, 2017); and a limited amount of training can significantly enhance performance in analytical tasks (Ares & Varela, 2017). These suggest that experienced, assessors who have been oriented to the sensory method could offer a viable alternative in situations where time, resources, or sample availability are constrained, especially in industrial applications (Giacalone & Hedelund, 2016; Barton et al., 2020; Wang et al., 2022).

There is growing consensus in the field that the choice between trained and untrained panels should be guided by the objective of the test (Meiselman, 2013; Ares & Varela, 2017; Barton et al., 2020; Maheeka et al., 2021), not by assumptions of superiority. When the goal is analytical, such as ingredient substitution, changes to formulation, or similarity testing, where smaller differences may be important

even if consumers might not notice them (Meilgaard et al., 2025), a trained panel is often more appropriate due to their heightened sensitivity.

However, for product development, consumer-relevant difference and consumer acceptability testing, where real-world usage and emotional responses matter, untrained or semi-trained consumer panels offer more relevant insights. While their responses may be less consistent, they more closely reflect how the average consumer experiences the product. This makes them especially valuable in early product development stages, and benchmark testing.

Even when trained panel data are available, there is still a need for additional tools to connect those results to actual consumer perceptions. Researchers note that data from trained panels often need to be matched with consumer liking data to fully understand product performance in the market (Ishii et al., 2007; Kemp et al., 2018). This remains an active area of research, highlighting an ongoing gap between technical sensory profiles and consumer relevance.

This ongoing debate about the appropriate sensory panel raises a number of questions: If trained assessors detect differences that consumers may not notice or care about, how meaningful is that added sensitivity in a typical use scenario (Ares & Varela, 2017)? If they function as analytical rating machines, do they belong more in fields like product engineering, given the aim of sensory evaluation is to capture subjective human perceptions, albeit in a structured and more objective manner (Meiselman, 2013)? And even if they are calibrated to use rating scales consistently, does that truly remove the influence of individual variability and scale-use bias?

The evidence suggests not. Moreover, individual variability persists across disciplines requiring human judgement, including sensory science (Næs, 1990; Romano et al., 2008; Sipos et al., 2021), psychometrics and educational assessment (Linacre, 1994; Myford & Wolfe, 2003; Engelhard & Wind, 2018), regardless of training or expertise. There will always be a need for techniques that can isolate true product differences, independent of the idiosyncratic use of scales by the assessors.

2.4.2 Rating scales: measurement, reliability and validity of results

In sensory analysis, various types of scales are used to convert subjective perceptions and associated sensory scores into measurable data. These scales

can be categorised into four basic types: nominal, ordinal, interval, and ratio (Stevens, 1946), and the type of scale chosen has significant implications for the validity of the measurements, the statistical methods used for analysis, and the interpretability of results (McEwan & Lyon, 2003; Meilgaard et al., 2025).

- **Nominal scales** classify data into distinct categories without any implied order. Common examples include binary responses such as yes or no, sweet or not sweet, or same or different, often used in paired comparisons and other difference tests. They also apply to classification tasks, such as sorting samples by colour or identifying the presence of an attribute, as in CATA questionnaires.
- **Ordinal scales** introduce a ranked order such as “weak”, “moderate”, and “strong”, and are sometimes used in attribute rating and hedonic testing. However, they do not convey the magnitude of difference between levels. While it is common practice to assign numerical values to these categories (e.g., a 1 - 9 hedonic scale where 1 signifies “dislike extremely” and 9 signifies “like extremely”) and treat the data as interval-level, this can be misleading, as it assumes equal spacing between categories, which often does not align with actual sensory perception (McEwan & Lyon, 2003; Næs et al., 2010; Boone, 2016).
- **Interval scales** place items into numbered groups separated by equal intervals, such as line scales. The numbers indicate both the order and a meaningful relative distance between values on the scale. However, they lack a true zero point (McEwan & Lyon, 2003; Bond et al., 2020), so statements about ratios (e.g., “twice as sweet”) are not valid.
- **Ratio scales** use numbers to express the magnitude of a stimulus as a multiple or factor of another. For example, indicating that a sample is twice as sweet as a reference sample. By including a true zero, which represents the absence of an attribute, ratio scales enable proportional comparisons. Magnitude estimation is an example of a ratio scale method (McEwan & Lyon, 2003; Meilgaard et al., 2015; Rogers, 2017).

Two main families of scaling methods dominate in sensory science and consumer research: line scales and category scales, each with advantages and limitations.

Line scales typically use a continuous visual analogue line, often 10 or 15 cm long, anchored with descriptors like "None" and "Very Intense" (Stone et al., 2012; Kemp et al., 2018; Meilgaard et al., 2025). Assessors indicate their perception by marking a point along the line, and the distance from the origin is measured and used for statistical analysis. These are widely used in QDA™ and Spectrum™ methods for their sensitivity and resolution, and results are often treated as interval-level for statistical purposes. Their continuous nature allows for a wide range of subtle responses, especially in trained panels.

However, their use can be cognitively demanding, particularly for untrained assessors, and they are sometimes prone to ceiling and floor effects. Ceiling effects occur when responses cluster at the top of the scale, limiting detection of improvements or distinctions among high-intensity stimuli, thereby reducing product differentiation. Conversely, floor effects occur when responses cluster at the bottom, obscuring subtle differences at low intensities. Both effects restrict the usable range of the scale, reduce variability, and compromise the sensitivity and interpretability of the data (Stone et al., 2012; Kemp et al., 2018).

Category scales simplify response collection by offering discrete options, typically in the form of hedonic or intensity categories anchored with verbal labels. The most familiar example is the 9-point hedonic scale, ranging from "Dislike Extremely" to "Like Extremely" (Peryam & Pilgrim, 1957). These scales are intuitive and widely used in large consumer tests for assessing preference and acceptability (Peryam & Pilgrim, 1957; Yao et al., 2003; Pham et al., 2008; Lesschaeve et al., 2012; Zhi et al., 2016; Ho, 2019).

Ordinal categorical scales are also used in overall difference testing (Schlossareck & Ross, 2019; Higgins & Hayes, 2020; Montero & Ross, 2022; Cela et al., 2023), as well as for attribute rating and descriptive analysis (Findlay et al., 2007; Reinbach et al., 2014; Puputti et al., 2019; Pineau et al., 2022). However, while convenient, these scales yield only ordinal data.

It is common practice, though methodologically debated, to treat ordinal responses as interval-level data for the purposes of statistical analysis, particularly using ANOVA. This approach assumes equal spacing between categories, normal distribution of residuals, and homogeneity of variance. These assumptions,

however, are not always met in response data (Næs, 1990; McEwan & Lyon, 2003; Ho, 2015; Boone, 2016; Raithatha & Rogers, 2018; Ho, 2019). While ANOVA is generally robust to moderate violations of its assumptions, particularly with larger sample sizes due to the central limit theorem (Kwak & Kim, 2017), the ordinal nature of categorical scales can present challenges when violations are severe (Stone et al., 2012) or sample sizes are small (Meilgaard et al., 2025). Specifically, residuals may not be normally distributed, and variances can be heterogeneous across groups due to individual differences in scale use, potentially compromising result validity. Most studies fail to report whether these assumptions were tested or met, raising concerns about the robustness of conclusions drawn.

Sensory data occasionally depart from normality because rating scales have fixed upper and lower limits (i.e., they are bounded), are ordinal in nature, and can be skewed by individual differences in scale use or reluctance to use extreme categories (Kemp et al., 2018). When data deviate substantially from normality, estimates of central tendency and variability may become biased (Stone et al., 2012; Meilgaard et al., 2025), reducing the sensitivity and interpretability of subsequent analyses. Understanding the distributional nature of sensory responses is therefore essential for selecting appropriate analytical techniques and ensuring that statistical conclusions accurately reflect perceived differences (Raithatha & Rogers, 2018).

Categorical-Ratio scales including the Labelled Magnitude Scale (LMS) (Green et al., 1993) and the Generalised Labelled Magnitude Scale (gLMS) (Bartoshuk et al., 2005) were developed to overcome limitations of both line and category scales. These are vertical, semi-logarithmic scales anchored with empirically spaced perceptual labels like “barely detectable”, “moderate”, and “strongest imaginable [the sensory stimulus being measured]”. Unlike linear or ordinal scales, the LMS aims to approximate ratio-level measurement by aligning verbal anchors with psychophysical intensity intervals derived and validated using ratio scaling (i.e., the magnitude estimation scale) (Lim et al., 2009). Bounded by “no sensation” and “strongest (or maximal) imaginable sensation” at each end, these scales enable comparison of individual and group differences within the full range of perceived intensities. Additionally, the inclusion of the verbal anchor “strongest imaginable”

was intended to minimise ceiling effects (Kemp et al., 2018), as assessors are instructed to rate sensations relative to the most intense version of the specific stimulus they can imagine (e.g., the strongest imaginable oral sensation). This personalisation broadens the scale's dynamic range and enhances discrimination among high-intensity experiences, reducing response clustering near the top and allowing for more accurate comparisons across individuals and stimuli.

The gLMS extends this approach by asking assessors to rate sensations relative to the strongest imaginable sensation of any kind, not limited to the same sensory modality. This adjustment was intended to reduce variability between individuals with differing sensory sensitivity levels, such as supertasters and non-tasters of bitterness (Bartoshuk, 1979), and to enable more meaningful cross-individual comparisons. However, the gLMS has faced criticism for assuming that individuals can reliably compare across sensory modalities. For example, a participant may be asked to rate the intensity of a bitter taste relative to the strongest imaginable sensation of any kind, such as the pain of a broken bone or the sound of a fire alarm. This type of cross-modal comparison can be cognitively demanding and may not be intuitive (Lim et al., 2009), especially when the sensations differ dramatically in both intensity and emotional relevance. Moreover, individual differences in prior experience, cultural background, and sensory exposure may influence how the upper anchor is interpreted, potentially reintroducing the very variability the scale was designed to minimise.

Both the LMS and gLMS scales have been criticised for their complexity and practical limitations, particularly when used by untrained assessors (Hayes et al., 2013). Common issues include the cognitive burden of interpreting their abstract anchors like "strongest imaginable", which requires conceptual effort and can lead to misuse or compression of the scale range. Scale bias arises when participants with limited exposure to high-intensity stimuli underuse the upper end of the scale, effectively narrowing the measured range (Schifferstein, 2012). There can also be considerable individual variability in how anchors are interpreted; what one assessor considers "very strong" or "moderate" can differ widely based on prior sensory experiences. Additionally, there is a tendency to use the scales as

categorical scales, with assessors relying solely on the semantic label anchors to assign ratings (Hayes et al., 2013).

Moreover, the semi-logarithmic and ordinal nature of LMS/gLMS data often violates key assumptions underlying parametric analyses such as ANOVA. Specifically, residuals may not be normally distributed, and variances can be heterogeneous across groups due to individual differences in scale use. These violations can compromise the validity of ANOVA results, highlighting the need for alternative approaches such as data transformations or non-parametric methods (Ho, 2015; Raithatha & Rogers, 2018) that better accommodate the unique properties and variability inherent in these semi-logarithmic scales and in ordinal rating scales. However, non-parametric tests like the Friedman (Friedman, 1937) and Kruskal-Wallis (Kruskal & Wallis, 1952) are rank-based and often considered a practical compromise, as they tend to reduce statistical power (Conover & Iman, 1981; Politi et al., 2021). Conversely, data transformations require an iterative and complex process, making them less feasible in many practical settings, especially in consumer or industry studies where time is constrained.

As Meiselman (2013) recommended, the choice of rating scales should be context-dependent, as there are no inherently good or bad scales. Instead, the focus should be on identifying the most user-friendly scale for the specific panel of assessors and the one most efficient in achieving the required results. However, regardless of the scale used, issues related to individual variability and response bias remain persistent challenges. Several studies have proposed methods to minimise and correct for the confounding influence of individual rating styles from true differences between samples (Næs, 1990; Romano et al., 2008; Brockhoff et al., 2015; Großmann et al., 2023), and (Sipos et al., 2025); however, these approaches generally address the issue at an aggregate data level. Working solely with averaged data can obscure important individual differences, masking individual rating tendencies and inflating measurement error.

Given these limitations in rating scale validity and individual variability, alternative approaches that model individual responses directly, rather than relying on aggregated data, offer promising solutions. Unlike traditional aggregation methods, a Rasch-based approach models the latent traits of both individuals and items,

enabling the disentangling of individual biases from true sensory differences. Specifically, the Many-Facet Rasch Model (MFRM) extends these principles to simultaneously account for multiple sources of variation including assessor severity, product differences, attribute characteristics, and other explanatory factors within a single measurement framework (Linacre, 1989). This approach enables the separation of person and item parameters, allowing for more precise measurement of sensory perceptions while accounting for variability in individual rating styles. The following section details the core principles and methodological extensions of Rasch measurement that position it as an effective framework for addressing these sensory evaluation challenges.

2.5 Rasch measurement

Rasch measurement is a psychometric approach used to measure latent traits i.e., unobservable characteristics or abilities (such as mathematical ability, user attitudes, or sensory sensitivities), by modelling the relationship between individuals and test items (i.e. survey or examination questions). Latent traits cannot be directly observed but are inferred through patterns in responses to carefully designed items or stimuli.

Developed by Danish mathematician Georg Rasch (Rasch, 1960), Rasch models use mathematical formulas to express the probability of a specific response (e.g., a correct answer or a sensory rating) as a logistic function of the difference between a person's latent trait level and the difficulty or intensity of an item on a linear scale (Lunz & Linacre, 1998; Boone et al., 2014; Ho, 2019; Bond et al., 2020).

Unlike traditional statistical models, such as regression or ANOVA, which fit a model to the observed data to explain patterns or differences, Rasch analysis operates by testing whether the data fit a predefined measurement model. Traditional tests require assumptions to be met, such as normality of residuals and homogeneity of variance, and typically rely on aggregated data. In contrast, Rasch analysis does not assume any specific underlying data distribution. Instead, it focuses on individual response patterns as the primary source of information (Wright, 1991; Linacre, 1999; Boone, 2016; Bond et al., 2020; Linacre, 2023b). When responses deviate from the model's predictions, these inconsistencies are flagged

for further investigation using built-in diagnostic tools based on residual analysis, such as fit statistics like outfit mean squares, which assess the response patterns of both persons (the respondents being examined) and items (the questions used in estimating the latent trait being measured), indicating how well they fit the model's expectations.

2.5.1 Types of Rasch Models

The original Rasch model was a basic dichotomous model and has since been extended by several researchers to address emerging research questions as its application became more widespread, as summarised in **Table 2.2.**

2.5.2 Key requirements and principles of the Rasch model

Rasch models specify several key criteria for a latent variable measurement to be meaningfully interpreted.

Unidimensionality: the core idea behind measuring latent variables is to draw inferences from observable data (what you have) to unobservable qualities (what you want but cannot measure directly) (Boone, 2016). For example, the questions in a math test should strictly measure mathematical knowledge, or perceived overall differences in flavour from a set of attribute intensity ratings should reflect the intended overall difference in flavour rather than unrelated factors.

Achieving unidimensionality is concept-dependent and empirically verified. It is a construct design decision, and so tests must be carefully designed to isolate and accurately capture the target latent variable by selecting theoretically aligned items and validating unidimensionality with the model (Smith, 2002; Linacre, 2023a, 2024a). If response patterns reveal that items measure multiple independent dimensions rather than contributing coherently to a single construct, the definition of the latent variable can be refined, or items split into separate analyses, either way, it provides valuable diagnostic insights to the researcher.

In sensory contexts where perception is often multidimensional and attributes often interact, the Rasch model does not claim to capture this full perceptual complexity. Rather, it measures whether a deliberately defined set of attributes works together coherently to reflect the researcher's intended construct, whether a

specific sensory modality (e.g., differences in flavour attributes) or an overall difference integrating multiple sensory dimensions (across taste, aroma, and texture). A construct can be unidimensional even when it includes cross-modal attributes, provided assessors use those attributes consistently to express the same underlying dimension. Researchers must therefore ensure unidimensionality through careful attribute selection and validation of response patterns.

Parameter separation: ensures that item difficulty (such as how bitter a sample is) is independent of the sample of respondents, and that individual sensitivity or ability is independent of the specific items tested. The model simultaneously estimates both individual sensitivity/trait levels and item difficulties directly from response patterns in the data, without requiring prior information about individual characteristics. This means that individual differences in sensitivity are accounted for. E.g., if a person consistently rates all samples as more bitter than other assessors, the model identifies this as higher sensitivity, independent of which samples were rated. This property, known as ***invariance***, allows for fair and consistent comparisons across different samples and assessors, enabling measurement that is both sample-free and item-free (Wright & Masters, 1982). In other words, a respondent's estimated ability does not depend on which items they answered, and item difficulties remain stable regardless of which respondents completed them, providing the foundation for objective and meaningful measurement.

Local item independence: as the model dictates that responses to each item depend only on the underlying latent variable, not on responses to other items. When items are more strongly related to each other than to the latent trait, they exhibit local item dependence (LID), which can bias measurement results (Sick, 2010).

Functioning of rating scale categories: Rasch analysis evaluates whether each response option on a scale is used consistently and in the intended order. If respondents struggle to distinguish between adjacent categories (e.g., confusion between "moderate" and "moderately strong"), the thresholds become disordered, signalling that the scale may need redesign or clearer definitions (Engelhard & Wind, 2018; Bond et al., 2020; Eckes, 2023).

Table 2.2. Summary of Rasch Models

Type of Rasch model	Use	Mathematical log-odds representation
Dichotomous Model (Rasch, 1960)	Used for binary responses. It estimates the probability that respondent n scores 1 instead of 0 on item i based on the difference between the respondent's ability (θ_n) and the item difficulty (δ_i).	$\ln\left[\frac{P_{ni}}{1-P_{ni}}\right] = \theta_n - \delta_i$
Rating Scale Model (Andrich, 1978)	The RSM is a polytomous model used when all items share the same response categories. It compares the probability of choosing category k to category $k-1$, with a threshold parameter (τ_k) representing the boundary between adjacent categories.	$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \delta_i - \tau_k$
Partial Credit Model (Masters, 1982)	The PCM is ideal for instruments where items have different scale structures, such as a mix of yes/no and rating questions. It handles varying numbers of response categories and allows each item to have unique step or threshold parameters (τ_{ik}).	$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \delta_i - \tau_{ik}$
Many-Facet Rasch Model (Linacre, 1989)	The MFRM extends the Rasch model to include multiple facets beyond persons and items, such as raters, samples, replicates, occasions, or other factors that could influence the responses. Where τ_k represents the threshold parameter, while the other symbols denote the various facets being modelled.	$\ln\left[\frac{P_{mnrik}}{P_{mnrik-1}}\right] = \beta_m - \theta_n - \rho_r - \delta_i - \tau_k$
Hybrid Rasch Models	These provide some flexibility to adapt the model to complex data, combining features of both the RSM and PCM models as described in (Linacre, 1994) and (Myford & Wolfe, 2003). For instance, in an MFRM, they can model how each rater m applies either a shared or a unique rating scale. Similarly, each item i can be evaluated using its own distinct scale structure. In these cases, the threshold parameter becomes τ_{mk} (rater-specific) or τ_{ik} (Item-specific) respectively, allowing the model to account for variations in scale use across raters or items.	$\ln\left[\frac{P_{mnrik}}{P_{mnrik-1}}\right] = \beta_m - \theta_n - \rho_r - \delta_i - \tau_{mk}$ or $\ln\left[\frac{P_{mnrik}}{P_{mnrik-1}}\right] = \beta_m - \theta_n - \rho_r - \delta_i - \tau_{ik}$

Non-linearity of ratings: Rasch models do not assume that response categories on a scale are equally spaced. Traditional methods such as ANOVA often treat ratings as interval-level data, implying equal psychological distances between categories, e.g., between “like moderately” and “like slightly”. However, this assumption can distort results, since category rating scales inherently provide ordinal data (Boone, 2016). This issue is particularly important in sensory research, where perceptions vary across individuals and scale intervals are unlikely to be uniform (Ho, 2019). Moreover, different items or attributes (e.g., sweetness vs. bitterness) may interact differently with the scale, producing unequal step patterns in the ratings.

Rasch analysis addresses this by using a probabilistic framework to transform ordinal responses into interval-level measures. This approach accounts for differences in both item difficulty (e.g., stimulus intensity) and respondent ability or trait level, enabling more valid comparisons across items and individuals (Wright & Masters, 1982; Linacre, 1994; Boone et al., 2014; Boone, 2016; Bond et al., 2020).

Reliability: Rasch analysis produces reliability indices that assess the consistency of measures across the latent trait continuum for persons (i.e. respondent), items, raters, and any other modelled variables. For example, if a sensory panel reliably distinguishes between mild and strong bitterness across samples, rater reliability will be high. Similarly, item reliability reflects how well the set of attributes spans the sensory continuum, ensuring adequate coverage of intensity levels. These metrics parallel classical reliability tests, but are grounded in the probabilistic Rasch framework (Wright & Masters, 1982; Linacre, 2023b).

Differential Item Functioning (DIF): Rasch models support the identification of differential functioning across various components of the measurement process, ensuring that results remain fair, interpretable, and reproducible across groups and testing conditions (Myford & Wolfe, 2004; Bond et al., 2020; Eckes, 2023). In sensory evaluation, this is especially important in global contexts where cultural background, language, and perceptual norms can influence how products are rated (Muñoz, 2002; Meiselman et al., 2022; Hort, 2024). DIF occurs when sensory attributes are interpreted differently by subgroups, such as assessors from different cultural backgrounds or with varying training levels, even when their underlying sensory acuity is comparable. These differences may reflect biases in perception,

scale interpretation, or prior experience rather than actual sensory differences. This makes DIF a valuable tool for supporting fairness and validity in diverse panels as well as in understanding drivers of product acceptability and satisfaction across cultures or target consumer groups. Rasch analysis can also detect **Differential Rater Functioning (DRF)** (Myford & Wolfe, 2004), which refers to systematic variations in how individual assessors use the rating scale. For example, some raters may consistently give higher or lower ratings to a specific group, which can suggest some bias towards that group affecting the reliability of results. This detection is especially important in longitudinal studies such as panel proficiency monitoring or repeated product evaluations (Tomic et al., 2010; Raithatha & Rogers, 2018), where systematic changes to an assessor's responses can be identified and further investigated. More broadly, Eckes (2023) groups these forms of differential functioning under **Differential Facet Functioning (DFF)**, which extends beyond items and raters to include other contextual factors such as time points, testing environments, or protocols.

Rasch analysis offers a powerful, diagnostic framework that simultaneously evaluates multiple critical aspects of measurement quality. This comprehensive approach provides researchers with a rapid yet thorough assessment of whether their data meet the rigorous requirements for valid and reliable measurement. By identifying responses that deviate from model expectations, disordered categories, local dependencies, and differential functioning across respondents, items, or contexts, the model ensures that observed differences truly reflect underlying sensory traits rather than artifacts of bias or inconsistency. Together, these principles establish a robust foundation for producing precise, fair, reproducible, and generalisable measurements of latent variables.

2.5.3 Current applications of the Rasch models

Several software packages are available for Rasch analysis (Rasch Measurement Transactions, 2025), with prominent options including **RUMM2030+** (Andrich, 1997-2025), **ConQuest** (Adams et al., 1997-2020), and **WINSTEPS®** with its many-facet version **FACETS©**, (Linacre, 2004a; Linacre, 2013; Linacre, 2025a, 2025e). Open-source alternatives exist in R, such as the ***eRm*** package (Mair et al., 2019) and ***TAM*** package (Robitzsch et al., 2021), expanding accessibility for researchers. Practical

guides for Rasch modelling are available in R Wind and Hua (2021) and Debelak et al. (2022). Despite difference in interface, algorithms, and terminology, these software packages implement core Rasch measurement models and provide relatively consistent assessments of data and measurement quality (Tennant & Conaghan, 2007; Ho, 2019).

Since it was first introduced in 1960 (Rasch, 1960), Rasch modelling has revolutionised objective measurement in social sciences. It was initially applied in educational assessments to measure constructs such as intelligence and mathematical ability. Today, its use extends to psychological testing, language assessment, medical and healthcare research, as well as consumer behaviour in business and marketing.

In **Education, and Language Assessment**, where Rasch analysis is most established (Eckes, 2023), it has been widely applied for the evaluation and validation of survey instruments and rating scales (Galli et al., 2008; Oon & and Fan, 2017; Samir & Tabatabaee-Yazdi, 2020); assess student and teacher performance (Zhang, 1996; Tavakol & and Dennick, 2013; Fan & Bond, 2019; Chi et al., 2021; Gordon et al., 2021; Quansah, 2022; Hiğde et al., 2024; Hariyono et al., 2025) and monitor rater behaviour (Engelhard Jr & Myford, 2003; Myford & Wolfe, 2009; Polat, 2020; Eckes, 2023). It has also supported bias detection through Differential Facet Functioning (DFF) and Differential Item Functioning (DIF) analyses. For example, Eskin (2023) applied DFF to identify native language bias in writing assessments, while Khalaf and Omara (2022) examined DIF across gender groups in an anxiety scale. Monitoring rater drift via Differential Rater Functioning (DRF) has been demonstrated by Myford and Wolfe (2009) and Eckes (2023), with recent extensions assessing differences between human and AI raters (Shin & Lee, 2024; Lamprianou, 2025). Comprehensive reviews and guidance on Rasch applications in education and language assessment contexts are available in (McNamara & Knoch, 2012; Aryadoust et al., 2021), underscoring the model's value in promoting fairness and validity in measurement.

In **Psychology**, Rasch models are widely used to measure latent traits such as anxiety, depression, and cognitive abilities. Freitas et al. (2014) validated the Montreal Cognitive Assessment Scale using Rasch analysis, while Dabb et al.

(2025) developed the Paternal Pregnancy-Related Anxiety Scale ensuring cross-continental relevance. Similarly, Adu et al. (2025) examined cross-cultural validity of the Depression Anxiety Stress Scales (DASS-21). In school psychology, Rasch models are often applied to assess student learning, behaviour, and rating scale performance. Boone and Noltemeyer (2017) provide practical guidance on the use of Rasch analysis in educational and school-based assessments.

Medical and Healthcare applications extensively utilise Rasch models to assess health-related quality of life and patient-reported outcomes, as well as to support the validation and cross-cultural adaptation of clinical and research questionnaires. Applications span fields such as rheumatology, nursing, physiotherapy, and pain management (Tennant et al., 2004; Taylor & McPherson, 2007; Tennant & Conaghan, 2007; Catley et al., 2013; Miller et al., 2016; Huang et al., 2018; Mohsen & Gill, 2019; Stolt et al., 2022; Tesio et al., 2024; Touzani et al., 2024; González-Pérez et al., 2025; Kim et al., 2025; Lu et al., 2025). Reviews by (Belvedere & de Morton, 2010) and (Christensen et al., 2024) highlight how Rasch analysis has evolved from a theoretical framework into a practical methodology, now routinely used to improve the accuracy and objectivity of patient assessments in both clinical care and medical research.

In **Business, Marketing and Consumer Behaviour Research**, Rasch models measure latent constructs like preferences, satisfaction, and brand perception. Early foundational work includes Bechtel (1985) who generalised Rasch models for consumer rating scales, and Lunz and Linacre (1998) who introduced multifaceted Rasch modelling for business and marketing applications. De Battisti et al. (2005) applied Rasch analysis to assess service quality perceptions amongst university students, while Pagani and Zanarotti (2010) applied it to analyse customer satisfaction data. Salzberger and Sinkovics (2006) utilised Rasch methods including DIF to detect bias across countries in international marketing data, and, Conejo et al. (2017) applied DIF to refine brand personality scales across demographic groups. Camargo and Henson (2015b) and Chalk (2020) used Rasch models to better align product features with user experience. More recently, Bassi et al. (2022) examined consumer responses to mountain product labels that indicated that product originated from mountain regions using Rasch analysis; Grispoldi et al.

(2023) validated scales measuring attitudes toward insect-based foods; and Prasetyaningrum et al. (2024) examined the impact of gamification on customer engagement in the banking sector. Collectively, these applications demonstrate how Rasch analysis continues to evolve beyond technical modelling to become an essential tool for enhancing the fairness, and interpretability of consumer-related measurement.

In the context of **Sensory Evaluation of Foods**, Rasch models have been applied to measure latent traits such as overall quality and overall liking by combining ratings of individual sensory attributes. They have also been used to evaluate assessor consistency and validate rating scales. Early studies laid foundational work for the use of Rasch models in this field. Garcia et al. (1996) demonstrated its utility in measuring sensory quality in Iberian ham as a latent trait derived from multiple sensory characteristics, showing that the Rasch model could successfully combine attributes such as flavour intensity, saltiness, and texture into a unidimensional quality scale. Alvarez and Blanco (2000) used the model to evaluate the reliability of olive oil tasting panels, finding that the Rasch model effectively identified inconsistent assessors and could improve panel reliability through targeted training. However, these applications remained largely isolated despite the methodological advantages Rasch modelling offers. Later studies, such as Andrés et al. (2004) on salt content and ham processing, and Bi et al. (2019) on aroma quality in hams treated with essential oils, cited the Rasch model for validating assessor consistency but failed to describe how the model was applied or what insights it provided, merely citing "García et al. (1996)" without further explanation.

Thompson (2003) applied Rasch scaling in wine judging to evaluate rater consistency and to refine both the sensory panel and the rating scales used, demonstrating that Rasch analysis could identify problematic rating categories and highlight judges whose ratings deviated systematically from the panel. Faye et al. (2013) focused on incorporating assessor expertise in wine glass sorting tasks and found that accounting for subject experience improved the interpretability of free-sorting data. A common issue across these earlier studies has been a lack of transparency and accessibility, with overly technical reports offering little guidance

on how Rasch modelling was implemented or why it provides advantages beyond traditional methods.

More recent studies have moved toward clearer and more accessible uses of Rasch models in sensory research. Ho (2019) introduced a multi-faceted model to measure overall liking based on several attribute ratings, arguing that single composite scores lack diagnostic depth. Mile et al. (2021) used a similar framework to study hedonic preferences for tilapia fish jerky, while Wu et al. (2021) explored the sensory impact of ginger-enriched pasta on both acceptability and satiety. Arboleda et al. (2021) developed perceptual scales for texture and refreshment in fruit juices, and Li's (2019) doctoral thesis investigated Rasch models for new product development and consumer research instrument refinement. Although Owusu et al. (2022) did not implement Rasch modelling, they proposed its future use for deriving composite liking scores in soymilk formulation research.

These more recent studies show that Rasch modelling can potentially improve the rigor, objectivity, and interpretive depth of sensory evaluation by accounting for assessor variability and uncovering the latent structure of sensory responses. The model offers several advantages demonstrated in sensory contexts: estimation of latent variables such as overall quality or liking from composite attribute ratings (Garcia et al., 1996; Ho, 2019; Arboleda et al., 2021; Mile et al., 2021), identification of inconsistent assessors and systematic bias patterns (Alvarez & Blanco, 2000; Thompson, 2003) and diagnostic identification of problematic rating categories (Thompson, 2003; Li, 2019; Wu et al., 2021).

However, compared to fields such as education, psychology, and healthcare, where the model is routinely used to address issues with rater bias and subjective scoring, its uptake in sensory evaluation remains relatively slow, even though sensory analysts routinely grapple with these very same challenges. This may be due to the technical complexity and lack of practical guidance on how to implement the model or apply its results. Ho (2019) and subsequent studies have begun to address this gap by demonstrating clearer, more accessible applications. Nonetheless, Rasch modelling is still rarely used in routine sensory evaluation practice. Sensory data continue to be analysed primarily by aggregating raw scores or means. While familiar and straightforward, it tends to mask assessor

inconsistencies, introduce scale-use bias, and offer limited diagnostic insight unless supplemented by several additional analyses.

Despite the methodological advantages of Rasch modelling, it remains underutilised in sensory evaluation contexts. This is likely because there is a lack of direct comparative studies demonstrating how the model performs against established methods such as traditional descriptive analysis or discrimination testing. Without such comparisons, practitioners may be reluctant to adopt it without clear evidence of practical benefits. Additionally, the psychometric and educational origins of Rasch measurement mean that existing guidance is often technical and lacks sensory-specific applications, creating barriers for practitioners trained primarily in traditional sensory methods. Addressing this gap through comparative studies and practical demonstrations tailored to sensory contexts is essential for advancing the adoption of Rasch modelling in sensory and consumer research.

2.6 Justification of study

This study explores and demonstrates the benefits of applying Rasch modelling to sensory difference testing. While existing research has used the Many-Facet Rasch Model (MFRM) to estimate latent variables such as overall sensory quality and overall liking based on combinations of sensory attributes, it has not yet been applied to quantify overall difference between products as a latent variable, which in turn can reveal which specific sensory attributes most influence perceived differences. This represents a missed opportunity, as current methods are typically limited in one of several ways: some analyse sensory attributes individually without integrating them into an overall difference score; others provide a single holistic measure without identifying the specific sensory attributes driving that difference; and some rely on qualitative insights without quantitative support or require complex, separate analyses to estimate overall difference. In contrast, modelling overall perceived difference as a latent variable within a Rasch framework offers a unified approach that provides both diagnostic clarity and quantitative rigour by combining holistic and attribute-level insights in a single interpretable analysis.

Moreover, sensory quality programs continue to struggle with individual differences in rating scale use. Existing statistical methods often fall short in adequately

accounting for individual rating tendencies and the inherently subjective nature of sensory data. These are precisely the kinds of challenges that Rasch modelling was designed to address and has effectively tackled for nearly seven decades in fields such as education, healthcare, and psychology, where human judgment is central.

This study presents a clear, step-by-step application of Rasch analysis in sensory difference testing, attempting to bridge the gap between the model's methodological strengths and its limited adoption in sensory evaluation. It highlights how a Rasch-based approach can improve data interpretation, reduce subjectivity, and support more consistent and actionable results. The Many-Facet Rasch Model (MFRM) is shown to be particularly useful for quality control and diagnostic analysis in contexts such as product development, ingredient substitution, benchmark tests, panel performance monitoring, and consumer research. The proposed method is especially beneficial to sensory analysts seeking faster, clearer, and more data-driven insights in a streamlined manner to support decisions about product differences.

Chapter 3

Rasch and General Analytical Methodology

3.1 Overview

This chapter provides an overview of the Rasch analysis and statistical procedures used across all three sub-study chapters.

3.1.1 A Rasch approach to sensory difference testing explained

Rasch analysis is a statistical method used to convert categorical data, such as surveys or rating responses, into interval-level measurements. Originally developed for educational assessments (Rasch, 1960), it is now widely applied across various disciplines that rely on human judgments.

Fundamentally, it allows researchers to estimate unobservable traits or latent variables such as mathematical ability, overall attitudes, or perceptions, based on patterns of responses to a set of observable items (e.g., exam or survey questions). The model estimates the probability of a given response as a function of the difference between the respondent's ability or trait level and the difficulty of the item. This approach places both item difficulties and respondent abilities on a common linear scale, converting ordinal raw scores into interval level measures, which supports more precise quantitative analysis (Boone et al., 2014; Bond et al., 2020).

In sensory evaluation, Rasch analysis can be adapted to address the challenges of subjective human ratings. Each sensory attribute is treated as an item (similar to questions in a survey), and each product or sample is considered the subject of measurement (similar to respondents in the model). Assessors often interpret and use rating scales differently, and these inconsistencies can obscure true differences between products (Raithatha & Rogers, 2018).

Rasch analysis addresses this issue by explicitly modelling and adjusting for such variability (discussed in section 2.5.2: **pg.34**). To measure differences between products, Rasch models estimate how each product scores on the underlying latent trait, in this case “Overall Difference”, based on the intensity ratings across

multiple attributes. The assessors are included as the rater/judge facet using the Many-Facet Rasch Model. The goal is to derive a fairer estimate of the overall sensory difference between products by accounting for variability in individual rating styles (e.g., scale level effects), rather than relying solely on aggregated averages. Averages can distort measurement results when assessors exhibit different rating effects or biases (Myford & Wolfe, 2003; Lawless & Heymann, 2010; Stone et al., 2012; Kemp et al., 2018; Sipos et al., 2021; Meilgaard et al., 2025). Aggregating scores without accounting for individual biases and differences in scale usage may lead to inaccurate representations of the true sensory characteristics of products.

3.1.2 The Many-Facet Rasch Model (MFRM)

The Many-Facet Rasch model (Linacre, 1989) extends the basic Rasch model (as shown in **Table 2.2. Summary of Rasch Models**) by allowing for the simultaneous analysis of multiple variables, referred to as **facets**, that represent additional sources of variation. In sensory testing, these facets may encompass combinations of the various variables including product samples, sensory attributes, order of presentation, panel groups, time of evaluation, replicate evaluations, and the assessors themselves, similar to parametric ANOVA methods.

Unlike traditional parametric approaches, which assume that all assessors interpret and use the rating scale in the same way, the MFRM explicitly models individual differences in rating behaviour by estimating a separate severity parameter for each assessor. These parameters reflect how strictly or leniently each assessor uses the scale compared to a neutral reference point. The model then adjusts the observed ratings based on these parameters using an iterative probabilistic process. Starting with initial parameter estimates, it calculates the likelihood of the observed ratings and repeatedly adjusts the severity parameters and other facets to maximise this likelihood. This fitting continues until the model converges on the best overall fit to the data, effectively calibrating all ratings onto a common scale (Linacre, 2023b). This adjustment allows for more accurate and fair comparisons across products by removing these systematic biases introduced by differences in individual rating tendencies (i.e., severe or lenient raters).

Even with comprehensive panel training, substantial variability remains in how individuals use rating scales, reflecting the influence of both stable individual variations such as genetic differences in sensory sensitivity (Bartoshuk et al., 2005), cultural background and prior experience (Brockhoff, 2011; Meilgaard et al., 2025), and transient conditions like fatigue, distraction, or mood during evaluation (Stone et al., 2012; Raithatha & Rogers, 2018). Thurstone (1927) showed that variability in human judgment can distort comparative evaluations, and subsequent measurement research has demonstrated how such rater effects can be systematically identified and, in the case of rater severity or leniency, statistically adjusted for using the MFRM (Myford & Wolfe, 2003). However, the model does not replace panel training but complements it by providing a diagnostic framework to detect and correct residual rater effects that persist despite training. By explicitly modelling assessor severity, it offers an approach to reducing extensive calibration sessions. Rather than attempting to enforce perfectly uniform scale use through training, it statistically adjusts for systematic individual differences in scale use (severity/leniency), thereby allowing training efforts to focus more on attribute understanding and discrimination. While training improves overall consistency, complete uniformity in scale use remains difficult to achieve in practice (Lawless & Heymann, 2010; Kemp et al., 2018).

While the ANOVA approach is generally robust to moderate violations of its assumptions due to the Central Limit Theorem especially with larger samples (Kwak & Kim, 2017), and can include assessors or replicates as fixed or random effects, it still treats differences among assessors as random noise rather than explicitly modelling them. In contrast, the MFRM treats these same factors as measurable facets estimated on a shared latent scale, allowing their direct comparison and providing individual-level diagnostics on rater severity and consistency within a unified probabilistic framework. This simultaneous estimation of product, attribute, assessor, and replicate parameters enables richer diagnostic insight and fairer comparisons than ANOVA alone.

When assessors use the scale consistently, ANOVA and MFRM may yield similar conclusions. However, when assessors differ systematically in scale use, for example when two assessors perceive the same sweetness level but one is more

expressive and routinely gives higher scores while another is more conservative, ANOVA does not separate these biases from true product effects. MFRM addresses this limitation by estimating and adjusting for individual severity parameters, thereby producing fairer product comparisons. While ANOVA remains appropriate when assessor variability is minimal or random, the MFRM provides an extension for cases where such effects are systematic and of diagnostic interest.

The MFRM results are often presented in a visual summary known as a **Wright map**, named after Benjamin D. Wright, a pioneer in Rasch measurement and educational assessment (Boone et al., 2014). In a sensory context, this map displays the relative positions of products, attributes, assessors, replicate evaluations, and any other modelled facets along a common latent continuum expressed in logits, providing a nuanced overview of the data structure.

In addition, the MFRM includes built-in diagnostic tools designed to evaluate the quality and integrity of the data. While the model accounts for individual differences in how assessors use rating scales, it still requires that their ratings remain internally consistent, as the estimation of all other facet parameters depends on these inputs. To evaluate this, MFRM provides several key diagnostics:

1. **Residual fit analysis:** the model identifies unexpected or inconsistent responses by flagging assessors who use the scale unreliably, and by detecting attributes whose ability to discriminate across products, assessors, and replicate evaluations (as relevant in this study) differs from that of other attributes in the facet (Linacre, 2012a; Wu & Adams, 2013; Eckes, 2023; Linacre, 2024b; and Linacre, 2025b).

2. **Rating scale category diagnostics:** detect when rating scale categories are used in a manner that deviates from the model's expectations, such as being underused or misunderstood. For example, if a seven-point scale is employed but certain categories are rarely selected, the model may suggest collapsing those categories. Unused or poorly defined categories may not contribute meaningful information, can confuse assessors, and may reduce measurement precision. Similarly, if categories are not clearly separated, it can lead to a lack of distinction between different intensity levels, compromising the effectiveness of the rating scale. In such cases, adjusting the scale can improve

measurement quality (Linacre, 2002b; Engelhard & Wind, 2018; Bond et al., 2020; Eckes, 2023). This is further discussed in section **3.3.1.4: Rating scale category diagnostics**.

3. **Principal Component Analysis of Residuals (PCAR):** detects systematic variation or correlation among items or attributes that are assumed to measure a single underlying latent variable (discussed further in section **3.3.1.3: Response dependency - Unidimensionality and Local Item Dependence (LID**). Rasch models assume that item responses are independent (i.e., the response to one item should not influence the response to another). When response dependency is observed and the rating on one attribute appears to determine the rating on another, it suggests that the attributes may be conceptually or perceptually related. This prompts further investigation into the nature of these relationships and their implications for the validity of the measurement (Tennant & Conaghan, 2007; Linacre, 2024a).

These diagnostics features will be discussed in more detail later in the chapter.

Overall, the Many-Facet Rasch Model (MFRM) offers a more transparent and nuanced approach to evaluating sensory data by modelling the process through which ratings are generated, rather than focusing solely on the final scores. This study aims to demonstrate how MFRM provides a complementary perspective by enabling analysis at the individual level and accounting for variation across multiple facets. It offers a practical and effective way to improve the reliability of sensory data and to gain deeper insight into the sources of variation within a test, without the need for extensive additional statistical analyses.

3.2 Framework for measuring Overall Difference using attribute intensity ratings

In this study, the MFRM was used to evaluate overall sensory differences between products based on the perceived intensity of multiple sensory attributes. Assessors rated each product on several attributes using ordinal categorical labelled scales. Ratings on these sensory attributes were collected across the products through sensory questionnaires. The goal was to combine these attribute ratings into a

single latent measure representing the overall difference between products, referred to as the ***Total Intensity Measure (TIM)***.

The basic assumption of this framework is that products with higher perceived intensity or more distinct attribute profiles are positioned higher on the Rasch logit scale. In this context, a product with a distinct attribute profile elicits strong responses across several sensory attributes that make it stand out relative to other products. This ease of differentiation by assessors can inform product development, positioning, or quality control decisions.

Sensory attributes are treated as items, each with its own difficulty parameter. Easier attributes tend to receive higher intensity ratings because they are easier to perceive, while harder attributes receive lower ratings due to being more difficult to detect. Products are treated as persons and are placed on the logit scale based on their combined attribute ratings. Assessors and repetitions are modelled as facets to account for differences in rating severity and variability across sessions, respectively.

The model estimates the probability of an assessor m, assigning a particular rating k to a product n, during a replicate evaluation r for a given attribute i, by considering several influencing factors. These include:

- The degree of leniency or severity (β_m) of an assessor (m) in assigning ratings.
- Total Intensity Measure (TIM) (θ_n) of the product (n), reflecting the overall difference and determining its location on the logit scale.
- The effect of the replicate evaluations (ρ_r) accounting for variation across repeated assessments (r).
- The degree of intensity (δ_i) of a sensory attribute (i), indicating how easily it was perceived across products.
- The thresholds (τ_k) between adjacent rating scale categories (k); for example, how much more intense an attribute must be to move the rating from “weak” to “moderate” intensity.

Mathematically, the probability of observing a rating in category k is modelled as a function of the relative distance between these facets:

$$\ln\left(P_{mnrik}/P_{mnrik-1}\right) = \beta_m - \theta_n - \rho_r - \delta_i - \tau_k$$

…Equation 3.1

This modelling approach allows for product comparisons that are adjusted for assessor rating behaviour and attribute difficulty, making the results more reliable than those from simple average scores. Unlike conventional methods such as ANOVA, which rely on assumptions about normal distribution and equal intervals between scale points, the Rasch model relies solely on response patterns in the data (Smith, 2002; Linacre, 2004b; Boone et al., 2014; Bond et al., 2020). Observations that do not fit the expected patterns are flagged and can be further examined. This provides a layer of quality control that traditional methods do not offer.

The Total Intensity Measures (TIM) generated by the model are then used to perform post hoc pairwise comparisons to identify significant differences between products. Since the measures are adjusted for assessor severity and attribute difficulty, they capture product differences more accurately than raw averages. Linacre (1989) explains that Rasch calibration places all facets on a common logit scale, enabling direct comparisons, while Myford and Wolfe (2004) noted that adjusting for individual rater severity improves the fairness and precision of comparative evaluations. In sensory data, lenient assessors' higher scores and strict assessors' lower scores are calibrated on the logit scale, ensuring that product differences reflect sensory variations devoid of their rating tendencies. This adjustment improves the quality of the data, making it more suitable for both parametric and non-parametric statistical tests. Boone et al. (2014) illustrate how Rasch-derived measures yield more valid interval-level estimates than raw mean scores, and Bond et al. (2020) noted that the resultant interval scaling and reduced bias better meet assumptions of parametric tests. Even when parametric assumptions remain unmet, non-parametric tests applied to Rasch measures gain increased sensitivity and accuracy because the calibrated data reduce uncontrolled variability and noise. Together, these advantages support better decision-making by identifying perceptual differences with minimal confounding effects from individual rating styles or other modelled sources of bias. While these advantages are well documented in psychometric and educational measurement,

their practical implications for sensory and consumer research remain largely unexplored.

The Rasch model also provides useful diagnostic tools, including:

- **Assessor fit statistics**: which indicate which assessors rated consistently and which ones deviated from the model's expected patterns for the panel.
- **Attribute fit statistics**: that identify which attributes contributed most or least to the overall difference latent variable.
- **Category diagnostics**: which reveal whether all parts of the rating scale function as intended, and
- **Principal Component Analysis of Residuals (PCAR):** which helps detect underlying sensory dimensions or interactions between attributes that might not be evident from conventional analysis.

This Rasch-based framework enhances traditional sensory analysis by providing clear, actionable insights. It adds depth and precision that support product development, innovation, panel management, and quality control in a more targeted and resource-efficient way, making it a valuable addition to existing sensory quality management methods.

3.2.1 Conceptualising Overall Difference as a latent variable

The content of this section is reproduced from Ariakpomu et al. (2025b)**.**

The theoretical development of measurement instruments for Rasch analysis requires careful design to accurately capture the parameters of the latent variable being measured (Boone, 2016; Engelhard & Wind, 2018). For this study, the construct modelling framework described by (Ho, 2019) was adapted to define *Overall Difference* as a latent variable, as presented in **Figure 3.1**. As previously discussed, this latent variable is estimated from assessors' intensity ratings of selected attributes. Within the Rasch framework, each attribute functions as an item defining the latent variable, each sample represents the respondent being assessed, and each assessor represents a rater with a unique severity level.

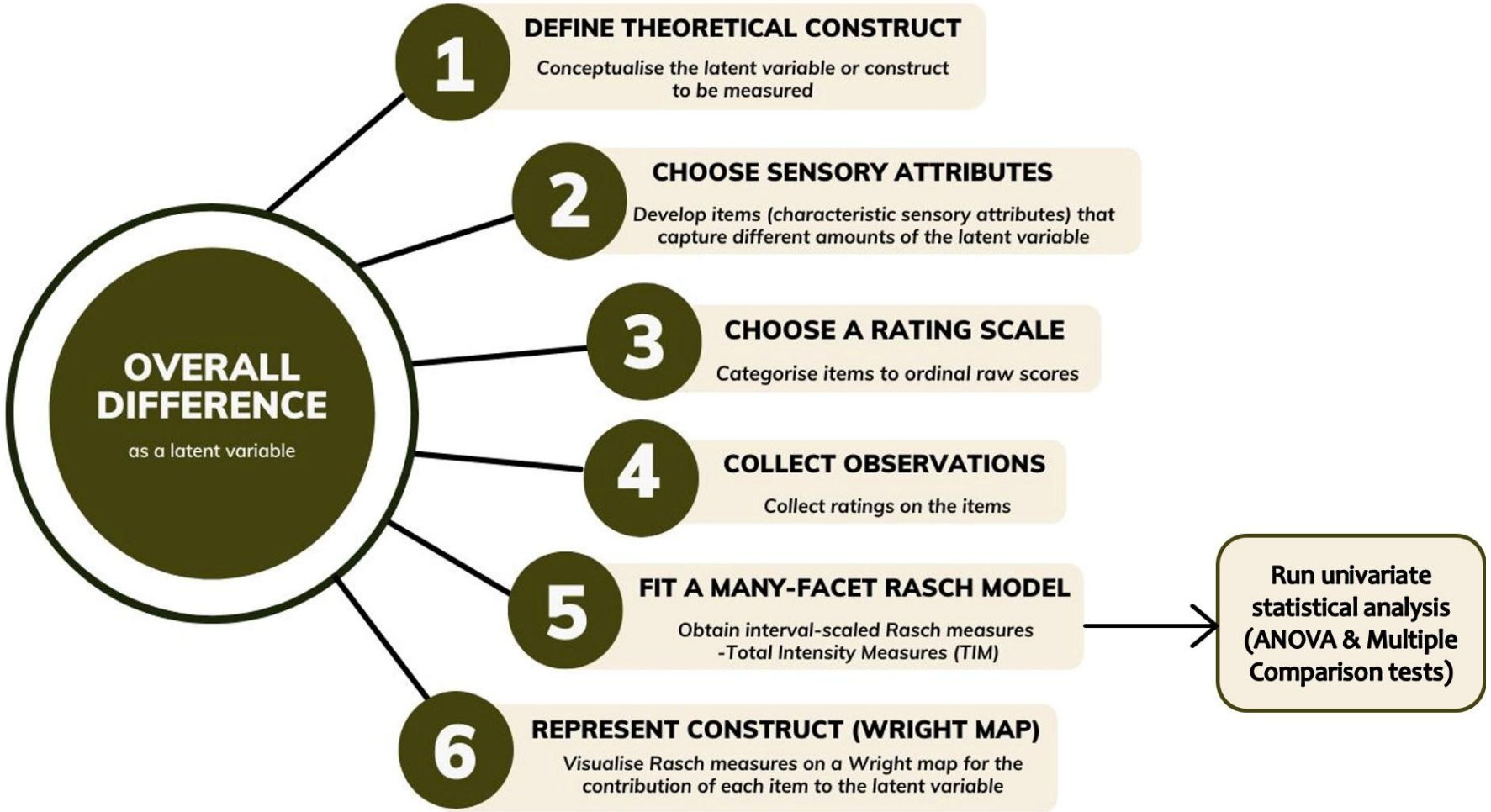


Figure 3.1. Framework for the conceptualisation of Overall Difference as a latent variable.

Step 1: Defining the theoretical construct of Overall Difference

Sensory attributes representing the sensory characteristics and modalities of the samples should be identified to capture different aspects of the *Overall Difference* latent variable. It is recommended to select a minimum of 3 to 5 sensory attributes to ensure sufficient variability in the data and allow the Rasch model to effectively separate the effects of different facets.

Step 2: Selection of attributes and survey design

Survey questions were developed for assessors to rate the perceived intensity of each attribute for each sample. For example: “How strong is the orange flavour for sample XXX?” These questions represent the items in the Rasch model.

Step 3: Choosing a rating scale

Labelled category rating scales (as described in the sensory testing procedures for the AR tests) representing levels of perceived intensity, were used by the panel of assessors - the raters for the Rasch model.

Step 4: Data collection

Observations were collected as attribute intensity ratings for each sample using the survey questionnaire developed in step 2.

Step 5: Fitting the Model

A Many-Facet Rasch Model (MFRM) with four facets - assessors (raters), samples (persons), attributes (items), and repetitions (replicate assessments), was fitted as described in ***...Equation 3.1***. The resulting Total Intensity Measures (TIM) for each sample were then used for post hoc multiple comparison tests to identify the significant differences between samples.

Step 6: Visual representation

The Rasch model’s Wright map visually represents the location estimates for each individual element within each facet (i.e., each assessor, product, attribute, and replicate assessment, referred to as **parameters** in this study), as well as the rating scale thresholds. These are all mapped on a common logit scale, providing a rapid overview of the underlying data structure and the relationships between the modelled variables.

This framework formed the basis for the data analysis applied across all three studies in this thesis.

3.3 Data analysis

All statistical analyses were conducted using RStudio version 2023.3.1.446, "Cherry Blossom" release (Posit Team, 2023), while Rasch analyses were performed using FACETS © version 4.3.0 (Linacre, 2025a) and WINSTEPS® version 5.9.0.0 (Linacre, 2025e).

3.3.1 Rasch analysis

Rasch analysis across all three studies was conducted using the Many-Facet Rasch Rating Scale Model (Andrich, 1978), as all attributes were assessed using a common rating scale.

3.3.1.1 Fitting the Many-Facet Rasch Model (MFRM)

The Many Facet Rasch Model (MFRM) simultaneously accounts for multiple variables, or facets, by modelling the log odds of observed ratings on a common interval scale known as the **logit scale**. Parameter estimates for the facets were obtained using Joint Maximum Likelihood Estimation (JMLE) in the FACETS software (Wright & Panchapakesan, 1969; Linacre, 2023b and ; 2025a). This method estimates all the facet parameters (i.e. assessor severity, product differences, effect of replicates, attributes intensity) at the same time, maximising the likelihood that the observed data fit the model. This joint estimation process continues until the model converges on the most probable set of facet locations on the logit scale.

The resulting parameter estimates for individual elements within each facet (i.e. facet parameters) were then visualised using a Wright map, providing a clear representation of their relative positions along the latent continuum. On the Wright map, the *Sample* facet was left non-centred, while the other three facets, *Assessor*, *Repetition*, and *Attribute*, were centred so that the mean of their parameters was zero. This centring established a common reference point on the Wright map, allowing the relative positions of samples to be interpreted in terms of the Total Intensity Measure (TIM). Consequently, sample locations were adjusted by

accounting for assessor severity, attribute intensity, and replicate session effects, corresponding to the *Assessor, Attribute*, and *Repetition facets*, respectively.

3.3.1.2 Global model fit

Rasch models are idealisations of empirical data, assuming that a single latent variable represents the underlying truth. For example, when data do not align with model expectations, they may distort this representation but can also reveal important issues such as disengaged students or flawed scoring rubrics, which are potential sources of measurement bias in educational contexts (Linacre, 2023b).

Assessing global model fit helps determine the practical usefulness of the data before further analysis. This involves evaluating whether the data fit the model in a meaningful way (Engelhard & Wind, 2018; Eckes, 2023; Linacre, 2023b), identifying the extent of any misfit, understanding its sources, and deciding how to address them. Model fit is typically assessed by comparing observed responses to those expected by the model, with differences usually expressed as standardised residuals. In this study, a satisfactory model fit is indicated when no more than 5% of absolute standardised residuals are ≥2, and no more than 1% is ≥3 as is recommended by (Linacre, 2022),

A meaningful fit means that, despite some imperfections in real data, the response patterns are still consistent enough with the model to support valid and interpretable measurement of the intended construct. It reflects a balance between the model's expectations and the complexity of real-world data. When misfit is observed, a closer inspection of the deviations from model expectations can reveal sources of bias. Based on these findings, model specifications can be adjusted, such as by removing an inconsistent rater, or combining overlapping items to improve overall fit.

3.3.1.3 Response dependency - Unidimensionality and Local Item Dependence (LID)

The Rasch model assumes that all items measure a single underlying trait. In this study, the **items** are the sensory attributes, and the trait is Overall Difference. To test this assumption, Principal Component Analysis of Residuals (PCAR) was conducted using WINSTEPS®, a Rasch measurement program designed for

rectangular data matrices with only two variables*. Following the procedure described by (Eckes, 2023), each attribute was placed in a column, and each combination of assessor and sample was placed in a row using the dialog box provided in FACETS.

PCAR was used to examine whether the unexplained variance in the residuals is small compared to the variance explained by the Rasch measures. If certain items exhibit similar unexpected patterns, this may indicate the presence of a second dimension. Whether or not this second dimension affects measurement depends on its strength. According to (Linacre, 2024a), a secondary dimension must have the strength of at least two items to be considered meaningful. If the variance in the residuals is large and attributable to a second dimension that the researcher deems significant enough to affect the interpretation or usefulness of the measures, remedial steps may include removing the responsible items or grouping items into subtests to define additional latent variables.

A related concern is Local Item Dependence (**LID**), where the response to one item can predict responses to another, contradicting the Rasch model's requirement for independent item responses (Tennant & Conaghan, 2007). LID is typically flagged when the correlation of standardised residuals between two items is greater than 0.3 (Ramp et al., 2009; Christensen et al., 2017).

However, the primary goal of PCAR is to detect these systematic patterns of co-variation. It is then the researcher's task to explore whether these patterns reflect meaningful conceptual differences or measurement bias, and to decide whether to retain, combine, or remove items (Smith, 2002; Hagell, 2014; Eckes, 2023; Linacre, 2024a).

In this study, the latent variable of interest was Overall Difference, based on a combination of sensory attributes across multiple modalities, and therefore inherently multidimensional. Signs of secondary dimensions or local dependency were expected and were not treated as sources of error, but as meaningful perceptual interaction between attributes. For this reason, no attributes were

* PCAR functionality is currently only available in WINSTEPS®.

removed or combined, as doing so would have resulted in the loss of valuable information about how sensory differences were perceived.

3.3.1.4 Rating scale category diagnostics

In Rasch analysis, scale category diagnostics evaluate whether the rating scale functions as intended, by examining how assessors use each category and whether this usage aligns with the model's assumptions. For the scale to function properly, categories should be used in a logical, ordered manner, with each one clearly representing a distinct level of the latent trait (e.g., intensity or difference) and receiving a sufficient number of responses. This helps determine how well the scale captures the latent trait and can inform improvements to both the scoring instrument and assessor training (Engelhard & Wind, 2018).

Guidelines recommended by Linacre (2002a); Engelhard and Wind (2018); Ho (2019); Bond et al. (2020); and Eckes (2023) for diagnosing the functioning of rating scales are summarised in **Table 3.1** below. These category diagnostics should be used in combination, as they typically tell the same story in different ways, and one often affects the other (Bond et al., 2020). For example, low category frequencies can cause disordered Andrich thresholds, resulting in probability curves without distinct peaks, reducing the precision and interpretability of the model's estimates (see **Appendix D**).

Some criteria are essential for evaluating the quality and measurement accuracy of the current dataset, while others are crucial when the scale is intended for use across multiple datasets, such as in developing new measurement instruments (Tennant et al., 2004; Galli et al., 2008; Conejo et al., 2017; Grispoldi et al., 2023; Dabb et al., 2025), exam rubrics (Tarricone & Cooper, 2014; Bond et al., 2020; Fidan et al., 2025), or rating scales for specific product categories in sensory quality programs (Thompson, 2003; Camargo & Henson, 2015a), where inference and generalisation are required.

Table 3.1. Guidelines for assessing the functionality of a rating scale.

Source (Linacre, 2002a; Engelhard & Wind, 2018; Ho, 2019; Bond et al., 2020; Eckes, 2023).

Criteria	Description	Implication
Item Polarity *(Essential for description of the samples[1], measure stability[2], measure accuracy[3] & inference[4])*	Scales should be positively oriented in the direction of the latent variable, so that higher ratings imply more of the latent variable. Point-Biserial (PT measure) correlation[5] for the item facet (attributes) should not reveal both negatively and positively orientated items (attributes).	A negative PT measure for an item suggests that items do not align with the theoretical expectation of how the latent variable should be measured. E.g., where higher item scores indicate less of the trait being measured. This often reflects confusion about the interpretation of the rating scale, i.e., do higher scores indicate more or less the attribute's intensity?
Category Frequency *(Essential for measure stability)*	There should be at least 10 observations in each scale category.	Category thresholds may be estimated poorly making it difficult for categories to describe distinct locations on the latent variable.
Observed Average Measures *(Essential for sample description, measure accuracy & inference)*	Computed as the average of the combined measure statistics of all the facets involved in producing scale category ratings. It should monotonically	Higher average measures will indicate ratings in higher scale categories and *vice versa*.

[1] Description of the sample refers to accurately summarising the observations in the study, i.e. how assessors perceived and rated sensory attributes across the samples.

[2] Measure stability refers to the consistency of a measurement system when repeated over time in the same context. E.g. the reproducibility of sensory ratings across different panels.

[3] Measure accuracy indicates how closely a measurement reflects the true value of the latent trait being assessed, i.e. how well categories and attributes differentiate between levels of overall difference

[4] Inference involves drawing conclusions about a broader population based on the sampled data and measurement results enabling generalisations beyond the current panel or samples.

[5] This measure is the MFRM equivalent of the Pearson point-biserial correlation (Linacre, 2023b). It assesses the relationship between responses to a specific item and the overall latent trait. A positive point-biserial indicates that the item aligns with the latent construct, while a negative value suggests misalignment, possibly due to item wording or misunderstanding.

	increase as the scale categories advance.	
Category model fit *(Essential for measure accuracy)*	Scale category outfit mean-squares indicate the deviation of average measure from the expected measures if data fit the Rasch model.	Category outfit mean-square statistics with values above 2.0 indicate that the category has been used in a different context than is expected.
Category Frequency Distribution	Frequency distribution of scale categories should be unimodal and tend towards a uniform distribution.	Intermittent low-frequency categories within the distribution may indicate irregular scale usage and the presence of redundant categories.
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 recommended distance between Rasch-Andrich thresholds is calculated[6] as 1.4, 1.1, 0.81, 0.70, 0.57, 0.51, and 0.45 logits for rating scales with 3, 4, 5, 6, 7, 8, and 9 categories, 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.

[6] Central distance =ln $(x/(m-x+1)$. For x =1,…, m, where m = n-1 for a n-category scale (Ho, 2019).

When indicators for the proper functioning of rating scales are unmet, remedial actions generally involve combining adjacent categories and sequentially renumbering the scale. In cases of item polarity, where some items are positively worded and others are negatively worded, it is important to ensure that items are reworded or properly reverse-coded to align with the theoretical expectations of the latent variable. A common example in sensory testing occurs when an attribute descriptor is not clearly defined, making higher and lower intensity ratings ambiguous. For instance, if assessors are unsure whether a higher rating means more or less of the attribute, their responses can become inconsistent. Failure to do address this issue is often flagged by a negative point-biserial correlation, indicating that the item may be misaligned with the construct and potentially misinterpreted by respondents. However, revising scale categories should not be undertaken without clear justification. As Linacre (2002a) notes, collapsing categories can reduce the precision and diagnostic value of the data, and should be approached with caution.

In this study, no category revisions were made, as the objective was not to optimise the rating scale for broader generalisability, but to examine how assessors utilised the existing scale structure. Retaining the original categories enabled a more accurate assessment of response patterns and scale functioning within the context of the current datasets. Revisions to the scale would have been necessary if the goal had been to adapt the scale for use with other samples of the same product or to enhance the measurement tools for broader application.

3.3.1.5 Separation statistics

Rasch separation statistics indicate how well a measurement instrument can distinguish between different levels of the latent variable across facets, such as persons (sample products), items (attributes), raters (assessors), and replicate sessions and how reliably those distinctions can be made. In other words, they show how effectively the scale differentiates between parameters in all modelled facets along the latent trait continuum (logit scale), as well as the consistency of these distinctions (Myford & Wolfe, 2004; Bond et al., 2020).

- **Fixed effect Chi-Square (χ^2):** This statistic, also referred to as the ***homogeneity index*** (Eckes, 2023) and reported as the ***fixed (all same) chi-square*** in FACETS (Linacre, 2023b), tests the null hypothesis that all elements within a given facet have the same measure after accounting for measurement error. In other words, it assesses whether all raters are equally severe or lenient, all attributes have the same intensity, samples differ significantly, or replicate evaluations are consistent. A significant fixed chi-square value ($p < 0.05$) indicates that at least two elements within the facet differ statistically (Myford & Wolfe, 2003; Eckes, 2023; Linacre, 2023b).
- **Separation ratio:** is a measure of the spread of the measures relative to their precision and is expressed as a ratio of the true variance to the error variance. Where true variance is the standard deviation after adjusting for measurement error (Myford & Wolfe, 2003; Linacre, 2023b). Higher values within a facet indicate better separation.
- **Strata:** refers to the number of distinct, measurable levels that a measurement instrument can differentiate along the latent trait continuum (represented on the logit scale), after accounting for measurement error (Myford & Wolfe, 2003). This measure is derived from the separation ratio* and is based on the assumption that the extreme ends of the trait distribution reflect meaningful and interpretable differences. According to Linacre (2023b), strata are appropriate when low or high values are interpreted as true differences, whereas separation is preferred if such extremes are considered to result from random variation. While separation indicates how widely measures are spread relative to measurement error, strata offer a more intuitive interpretation by representing the number of distinct levels or bands that the measurement can reliably differentiate within a facet.

In this study, strata were reported both for ease of interpretation and for consistent analysis across facets, rather than using a mix of strata and separation indices. This approach was intended to support uniform reporting across the modelled facets and maintain methodological coherence, while allowing for meaningful interpretation of

* Strata = (4 × Separation +1) / 3.

observed variation along the latent trait.

- **Reliability (Separation reliability):** is the MFRM's equivalent of the Cronbach's alpha test reliability statistic (Linacre, 2023b). It indicates how confidently the measurement tool can distinguish between elements within a facet. It is calculated as the ratio of true variance to total observed variance. A higher reliability value means that the ordering of measures (e.g. which sample scored higher, or which assessor was more lenient) is likely to remain stable if the assessment were repeated, suggesting that observed differences reflect real variation rather than random error. Values below 0.5 indicate poor reliability, suggesting that most of the variation is due to measurement error rather than true differences (Wright & Masters, 2002; Myford & Wolfe, 2003; Linacre, 2023b).

Facet reports both population and sample standard deviations (SD). Population S.D. is used when the dataset represents the entire population of interest, reflecting the true variability within that group. Sample S.D. is applied when the data are considered a subset drawn from a larger population, supporting generalisation beyond the group (Linacre, 2023b). In this study, population S.D. were used for each dataset because the focus was on variability within the specific assessors and samples studied, with no intention to generalise findings beyond them.

3.3.1.6 Residual fit statistics

"Fit is at the core of Rasch measurement" (Bond et al., 2020, p. 54). Fit statistics are fundamental to Rasch analysis, guiding the refinement of measurement instruments by identifying discrepancies between observed responses and the model's expectations, known as **residuals**. Residual fit statistics play a central role in evaluating data quality and underpin the diagnostic depth of MFRM by providing fit indicators for each element in every modelled facet (e.g., each assessor, attribute, or samples). Misfit arises when observed response patterns deviate from what the Rasch model predicts.

The two primary fit statistics used in WINSTEPS and FACETS are INFIT and OUTFIT. INFIT is information-weighted and more sensitive to unexpected responses near the predicted measure for an element, while OUTFIT is unweighted and more sensitive to outliers or extreme responses far from the expected values

(Smith, 2002; Eckes, 2023; Linacre, 2025b).

The unstandardised form of fit statistics, known as **mean squares**, represents the mean of the squared residuals (Bond et al., 2020; Linacre, 2025b). Larger residuals indicate greater misalignment between model expectations and observed ratings. The standardised form, expressed as a Z-statistic, adjusts for sample size and reflects how likely the observed level of misfit is to occur by chance under the model (Bond et al., 2020; Eckes, 2023; Linacre, 2025b). However, they are more sensitive to sample size and less informative about the practical magnitude of misfit as they only reflect whether misfit is statistically significant, but not whether it is large enough to matter for the measurement process.

In this study, unstandardised outfit mean square (OUTFIT Mnsq) statistics were selected for assessing fit. This choice was based on the nature of sensory data, especially from untrained panels, where extreme or inconsistent ratings are more likely to occur. OUTFIT statistics are more sensitive to these unexpected values than infit statistics, enhancing the ability to detect anomalies and assess measurement quality in detail. Furthermore, outfit mean squares are already adjusted for sample size as they are chi square statistics divided by their degrees of freedom, thus indicating the magnitude of the misfit rather than its probability of occurring (Linacre, 2025b).

Mean square values have an expected value of 1.0. Values significantly below 1.0 suggest **overfit** where responses are too predictable and contribute little additional information, often indicating redundancy and poor discrimination among variables, while values significantly above 1.0 suggest **underfit** or unmodelled noise, meaning responses are more erratic than expected. Values greater than 2.0 may indicate responses that distort the measurement. Although a commonly accepted fit range of mean square values considered “productive for measurement” is 0.5-1.5 (Linacre, 2025b), acceptable limits can vary depending on the context and sample size. This is because the variance of mean square statistics is inversely related to sample size (i.e., asymptotic variance = 2/Nr), so smaller datasets produce wider fluctuations around 1.0 (Wu & Adams, 2013). Consequently, fit ranges should be tailored to the assessment context (Bond et al., 2020; Eckes, 2023), and some researchers suggest using tighter ranges for high-stakes decisions and more

relaxed ones for exploratory or low-stakes assessments (Engelhard & Wind, 2018; Linacre, 2025b).

To calculate sample-size-adjusted fit ranges, Wu and Adams (2013) and Eckes (2023) recommended the formula shown in ***...Equation 3.2***, which yields wider acceptable ranges for small Nr and narrower ranges for large Nr, thereby improving the precision of fit diagnostics.

$$\textbf{Acceptable fit range} = 1 \pm 2\sqrt{\frac{2}{Nr}}$$

...Equation 3.2

Where **Nr** is the number of responses contributing to the parameter estimate within the facet of interest. For example, in the *Assessor facet*, Nr is the total number of ratings assigned by an assessor; in the *Attribute facet*, it is the total number of assessor ratings on that attribute. This formula was applied across all three studies to evaluate the performance of individual assessors and the contribution of attributes to the overall difference.

In this study, the results from the Rasch analysis were used primarily for diagnostic purposes rather than to develop a new measurement scale or refine an existing one, tailored to a specific set of items or products. Accordingly, no remedial actions such as collapsing rating scale categories or modifying item structures were taken, as these are typically part of an iterative development process (Engelhard & Wind, 2018; Tesio et al., 2024). Instead, the focus was on uncovering nuanced insights into the data and evaluating the performance of the modelled facets in terms of their consistency, interrelationships, and overall contribution to measuring the *Overall Difference* between products.

This diagnostic approach aligns with the perspective of Tesio et al. (2024) who emphasise that the Rasch model is not meant to *"transform messy data"* but to prompt researchers to reflect on the underlying causes of model deviations (Linacre, 1989; Linacre, 1994, 2023b) and iterate from there. Consequently, this study adopts a diagnostic stance in applying the Many-Facet Rasch Model (MFRM) to sensory difference testing.

3.3.2 Statistical analysis

Data preparation: No additional preprocessing or data transformation was applied to the sensory ratings. Raw assessor scores were entered directly into the respective analysis software without modification. For the ANOVA-based analyses, the ordinal ratings were analysed as recorded, while for the Rasch analyses, the same raw ratings were input into the FACETS program for parameter estimation. The resulting Rasch measures were then subjected to Kruskal–Wallis multiple comparison tests to evaluate overall product differences.

3.3.2.1 Product comparisons for Overall Difference

Statistical analyses for product comparisons were conducted on both raw score data and Rasch-transformed measures across all datasets to enable comparison of results of the two approaches. Differences between sample products were evaluated using both parametric and non-parametric analysis of variance (ANOVA) methods.

- **Parametric ANOVA:** ANOVA models were fitted using the R packages *MASS* (Venables & Ripley, 2002), and *car* (Fox & Weisberg, 2011). Residual analysis was performed with *nortest* (Gross & Ligges, 2015), and post hoc Tukey's HSD (Tukey, 1949) test for pairwise comparisons were performed with the *multcomp* package (Hothorn et al., 2008).

- **Non-parametric ANOVA:** through the *kwManyOneDunnTest* function, the Kruskal-Wallis test (Kruskal & Wallis, 1952), and the Dunn's Many-to-One Rank Comparison test (Dunn, 1964) for pairwise comparisons with a control were implemented using the *PMCMRplus* package (Pohlert, 2023). The Friedman test (Friedman, 1937), along with pairwise comparisons against a control using Nemenyi's Many-to-One Test for Unreplicated Blocked Data (Hollander et al., 2014) (via the *frdManyOneNemenyiTest* function), was also performed with this package.

  A Benjamini-Horchberg (BH) p-adjustment (Benjamini & Hochberg, 1995) was applied to control the false discovery rate, rather than the more conservative Bonferroni correction (Bonferroni, 1936), which controls the familywise error rates. The BH adjustment was preferred because it maintains greater statistical power and reduces the risk of Type II errors (i.e., failing to detect real differences

when they exist). This balance is important with Rasch measures of latent traits because these measures estimate subtle underlying constructs, and overly strict corrections can mask real differences.

For the Dunn many-to-one comparisons, a one-tailed alternative hypothesis ("greater") was specified for the DFC test results. This was justified because the pairwise comparison involved the DFC of the blind control (expected to show no difference or less effect as a placebo) and the DFC of the test samples, where differences were expected to be greater than those of the blind control. Conversely, for the Total Intensity (Rasch) Measures (TIM), a two-tailed alternative hypothesis ("two-sided") was used, since differences between samples and the control could be either an increase or a decrease in intensity.

3.3.2.2 Panel and assessor performance

Panel and individual assessor performance were examined trained and untrained panels by ANOVA-based methods in accordance with the performance criteria outlined in ISO 11132:2021 (British Standards Institution, 2021). The previously mentioned statistical packages were also employed in this analysis. Detailed descriptions of the analytical procedures are provided in the relevant chapters.

To investigate response patterns as recommended by (Stone et al., 2012; Ho, 2015; Raithatha & Rogers, 2018):

- **Response distribution plots (Trellis or lattice plots)** were used to evaluate the scoring behaviour of individual assessors across samples and replicate evaluations.
- **Assessor by Sample interaction plots** were used to investigate the agreement among assessors in the panel by plotting each assessor's mean responses relative to the panel average.

3.3.2.3 Data visualisation

Attribute contribution plots were created using Microsoft Excel 365 (Microsoft Corporation, 2019). All other data visualisations, including convergence, interaction and response distribution plots, were generated using *ggplot2* (Wickham, 2016) in R.

3.4 Data Collection

Quantitative data for this research were collected through sensory evaluation studies. All sensory questionnaires were developed and administered using RedJade sensory software (Redjade Software Solutions, 2023).

All three sensory studies were conducted in individual booths under white light at the Sensory Laboratory of the School of Food Science and Nutrition, University of Leeds, except for the trained panel data in Chapter 5, which were collected at a global chocolate manufacturing company's sensory testing facility in the UK.

The next three chapters provide a detailed discussion of the research themes under which Rasch analysis has been applied in sensory difference testing and quality control. As sensory evaluation methodologies varied across the three studies, each chapter includes a dedicated sensory methodology section. A brief overview is provided below.

Chapter 4 examines how Rasch analysis can be used to measure overall product differences by combining multiple attribute intensity ratings. The study used Difference from Control (DFC) and attribute intensity ratings from an untrained panel (n = 67) on three Jaffa cake samples. Attributes were selected based on an existing dataset whose corresponding study is discussed in Chapter 5.

Chapter 5 focuses on examining assessor performance with the Rasch model and compares the performance of a trained (n=7) and untrained panel (n=24) using three chocolate spread samples. Attributes were selected by the trained panel following a training phase.

Chapter 6 uses data from DFC and attribute intensity ratings from an untrained panel (n=54) on three tomato soup samples to explore the application of the Many-Facet Rasch Model (MFRM) as a unified approach for sensory quality programmes. Prior to the evaluation sessions, a preliminary session with untrained assessors (n=7) was conducted to generate sensory descriptors. From this, eighteen attributes were selected based on how frequently terms were mentioned.

Chapter 4

Measuring Overall Difference with the Many-Facet Rasch Model (MFRM): The Total Intensity Measure (TIM) Method

4.1 Overview

As part of routine quality assurance (QA) and quality control (QC), as well as in market research and product development, products are evaluated to identify differences between samples. The choice of a sensory test depends on whether the objective is to determine overall differences between samples or differences in specific attributes. To assess overall differences, the Difference from Control (DFC) (Aust et al., 1985) is quite beneficial as it evaluates the magnitude of differences between samples relative to a chosen standard, rather than just identifying whether differences exist (Whelan, 2017; Compusense, 2020; Montero & Ross, 2022). When the objective is to identify differences in specific attributes, methods like paired comparison tests and alternative forced-choice tests are used. These tests focus on one attribute at a time. However, sensory QC/QA often requires insights into multiple attribute differences between samples, which these methods do not efficiently provide. To address this, sensory descriptive methods are used to obtain intensity ratings for several attributes. Multivariate data analysis techniques like the Principal Component Analysis (PCA), are commonly used to interpret the data. PCA helps reveal patterns in the underlying data by reducing multiple attributes into fewer dimensions, providing a more comprehensive understanding of sample differences. However, interpreting results from such methods can be complex.

A Rasch approach could serve as an efficient alternative for measuring product differences offering both qualitative and quantifiable insights. As outlined in section **3.2: Framework for measuring Overall Difference using attribute intensity ratings**, the Many-Facet Rasch Model (MFRM) estimates a holistic Total Intensity Measure (TIM) for each sample, by combining attribute intensity ratings. TIM results are then subjected to univariate multiple comparison tests to quantify the overall difference between samples. Additionally, inherent Rasch quality control statistics provide deeper, easily interpretable insights, identifying which attributes were more

challenging for assessors to evaluate and determining the relative contribution of each attribute to the overall difference. This enhances diagnostic information, supporting more informed decision-making in sensory quality programmes with fewer tests.

This chapter compares overall difference measurement using two approaches: the traditional DFC test and the Rasch-based multi-attribute (TIM) approach, and forms part of the published article (Ariakpomu et al., 2025b).

4.1.1 Objectives

The aim of this study was to determine whether the TIM approach to measuring product differences is equally as effective as the DFC. The hypothesis is that the TIM method would yield similar overall difference results to those from the DFC, while the MFRM will provide additional insights on how individual attributes contribute to the *Overall Difference* construct.

The specific objectives were:

1. To evaluate the overall difference between three Jaffa cake samples using the DFC test.
2. To assess the intensities of five sensory attributes in the three Jaffa cake samples with the Attribute Rating (AR) test
3. To estimate the Total Intensity Measures (TIM) by combining the intensity ratings from the five attributes using the MFRM.
4. To compare the overall difference results from the DFC ratings and the TIM from the combined attributes using multiple comparison tests.
5. To interpret the additional insights provided by the MFRM's quality control statistics.

4.1.2 Study highlights

- TIM could differentiate between all three Jaffa cake samples while DFC could only differentiate between one of the samples and the control.
- The MFRM Wright map illustrated which attributes were easier and more challenging for the panel to perceive.

- Outfit mean square statistics for the attributes, combined with attribute logit values, revealed that *Orange flavour* had the highest contribution to the sample differences while *Saltiness* was the most challenging attribute for the panel to evaluate.

4.2 Sensory study: materials and methods

Sensory data were from the dataset referenced here as (Ariakpomu et al., 2024).

4.2.1 Samples

Jaffa cakes were chosen for this study as they share similar taste and flavour attributes[1] (*Orange flavour*, *Sweetness, Cocoa flavour*, *Milky flavour* and *Saltiness*) with the chocolate spread samples used in a related study (discussed in **Chapter 5**). They were selected to extend the MFRM validation to a more complex food matrix while maintaining experimental control. Jaffa cakes are sponge cakes with three layers: a sponge base, an orange-flavoured jam layer, and a chocolate top coating covering the side with the jam layer. This provided a heterogeneous food matrix with similar flavour characteristics to the chocolate spreads. The specific brands used in this study were chosen for their relatively uniform appearance, which was important for isolating taste and flavour differences from visual cues during sensory evaluation. Alternative chocolate-orange products, such as cookies or bars, were not selected because of their variable appearance across pieces and brands, which could introduce unwanted visual biases. Using products that share similar sensory attributes across both studies allowed examination of whether MFRM performs consistently across different food matrices when evaluating comparable sensory dimensions.

To facilitate comparability of the AR test with the DFC, efforts were made to ensure that all other sensory characteristics except taste/flavour were consistent across the samples to be tested. This was necessary because the DFC test only assesses overall product differences, meaning that attributes not included in the AR tests

[1] *Step one* (in Figure 3.1:Framework): conceptualise the latent variable by identifying sensory attributes to capture the overall difference dimension from the samples.

could still be perceived in the DFC, potentially influencing the conclusions drawn from the comparison.

The three selected samples comprised of one premium brand and two store-brand Jaffa cakes. These were chosen based on informal tasting sessions within the research team, and information from their back-of-pack labels. They were purchased from major supermarkets in the United Kingdom and differed in their nutritional and ingredient composition (**Table B 1**). The store brands were very similar in appearance, and one was selected as the reference for the DFC test, while the premium brand had a slightly different shape (**Figure B 1**). However, a significant limitation of this sample selection approach is that no instrumental analysis was conducted to verify that samples differed only in the target taste/flavour characteristics. While sample selection was guided by label information and visual inspection confirmed general uniformity in appearance, differences in texture properties (e.g., sponge density, jam consistency, chocolate coating hardness) and other non-target sensory attributes could not be ruled out and may have confounded the interpretation of the five focal taste/flavour differences.

The samples were stored in odour-free, airtight, plastic containers at room temperature (20±3˚C) until they were ready to be presented.

4.2.2 Participants

Ethical approval for the sensory study was granted by the Business, Environment and Social Sciences Faculty Research Ethics Committee at the University of Leeds.

Participants (n=67) were residents of Leeds, the majority of whom were staff and students at the University of Leeds. They were recruited through, emails, poster adverts and personal referrals and were selected based on the following criteria:

- Aged between 18 and 65 years
- Not having any chronic health conditions
- Not allergic or intolerant to the ingredients in the Jaffa cake samples
- Not on any routine medication (except contraceptives)
- Not on any special or restricted diets
- Not pregnant or lactating

- Available to attend two 1-hour-long sensory test sessions, within one month and with a minimum of four days between sessions.

Each participant was provided with detailed information about the study requirements, as well as the data protection and sharing disclaimer. They were then required to give informed consent by signing consent forms, both at the point of enrolment and a hard copy when they attended their first study session, to ensure they understood all study requirements and were happy to proceed.

The final untrained panel consisted of 43 females (64%) and 24 males (36%), aged between 18 and 54 years. They represented various ethnicities: 28 Asian (42%), 16 Black (24%), 15 White (22%), 2 Mixed (3%), and 6 from other ethnic groups (9%). To encourage commitment, participants were able to select two convenient dates for the sensory tests through an online form (Jotform Inc, 2023). The form was programmed to automatically send reminder emails 1 day, 2 hours, and 1 hour before their scheduled appointment. After completing the two required sessions, each assessor received a £20 Amazon voucher as incentive for their participation.

Ideally, trained panels with fewer assessors are used in QC settings. However, this study employed a larger number of untrained assessors to explore the TIM approach across varying levels of sensory expertise.

4.2.3 Study design

A Randomised Complete Block Design (RCBD) and Latin Square (**Figure 4.1**) were used to account for order effects and other sources of variation in the sensory experiments.

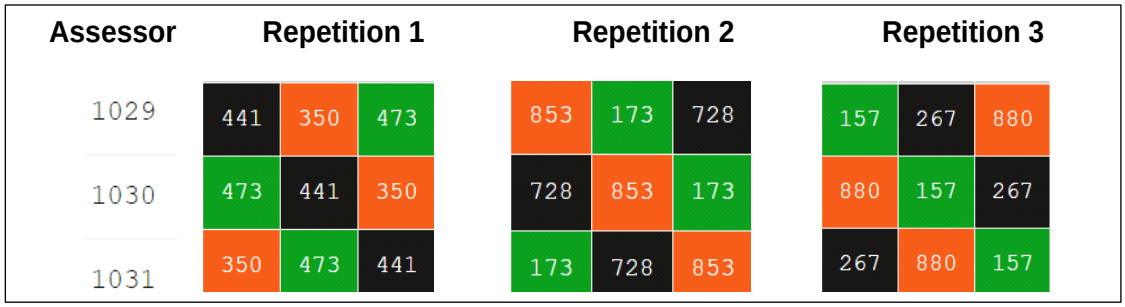


Figure 4.1. Illustration of sample presentation order in a Latin Square, showing three assessors (1029, 1030, and 1031) for the three samples (represented by different colours) across three replicated sessions. Source (**Redjade Software Solutions, 2023**).

In this design, each assessor evaluated three samples in a different order across replicate sessions, with each sample appearing in every position and following every other sample an equal number of times (Næs et al., 2010).

Each assessor participated in two separate sessions, one for the DFC test and another for the AR test, with a minimum interval of four days between each test session. To minimise expectation biases (Meilgaard et al., 2015), half of the participants completed the AR test first, while the other half started with the DFC test. Additionally, to reduce experimental variations due to the time of day, participants could only choose two morning sessions or two afternoon sessions for both tests, with the appointment booking form programmed to automatically account for this. Attendance was balanced according to both the time of day and which of the two tests the participants first completed.

In each test 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 nine evaluations for AR and eighteen for the DFC. All samples were served at room temperature (20±3°C) on 15 cm white paper plates labelled with random 3-digit codes. The reference sample for the DFC was labelled “R”.

4.2.4 Sensory evaluation procedures

The Difference-from-control (DFC) test followed the procedure described by (Meilgaard et al., 2015). 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-flavoured jam centrally positioned on one side of the sponge base, which is then covered with a layer of chocolate (see **Figure 4.2**). Without this guidance, assessors might only take a bite from the edge, missing the orange-flavoured centre and compromising the uniformity of the sample evaluation.

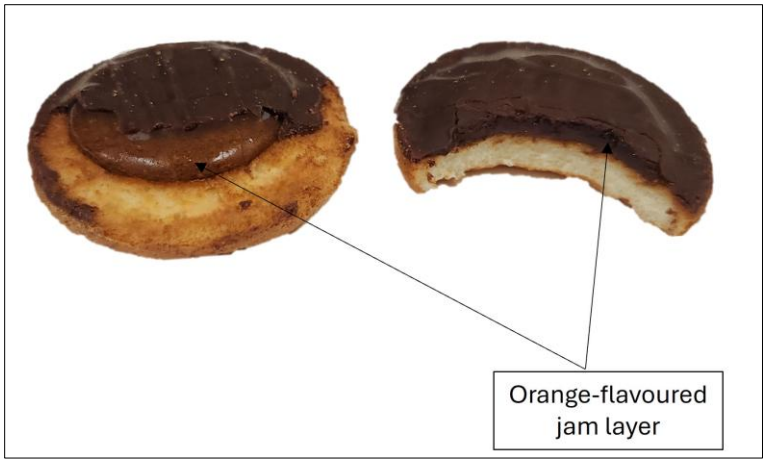


Figure 4.2. Photo showing the side of a Jaffa cake with centrally located orange-flavoured jam layer on the sponge base (with a portion of the chocolate top coating removed) justifying why assessors were instructed to take semi-circle-shaped half bites.

They were instructed to first taste the sample labelled "R", then taste the coded test sample, assess the overall difference between them and then rate the size of difference perceived. Assessors used a unidirectional labelled 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, [2]assessors rated the perceived intensities of five taste/flavour attributes: *Orange flavour*, *Sweetness*, *Cocoa flavour*, *Milky flavour*, and *Saltiness*. As previously mentioned (**4.2.1 Samples**), these attributes were selected based on a preliminary study involving products with similar taste/flavour characteristics, where a trained panel from a global chocolate manufacturing company identified these attributes for orange-flavoured chocolate spreads. The same attributes were used in this study to explore the Rasch-based 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 randomised for each sample and assessor, as suggested by (Ares et al., 2014) attempting to reduce errors of habituation, logic

[2] *Step two* (in Figure 3.1:Framework): design questions based on selected sensory attributes to capture different amounts of the latent variable.

and halo effect (Lawless & Heymann, 2010). [3-4]An 8-point categorical intensity scale ranging from 0-7 with labels adapted from the Labelled Magnitude Scale (LMS) (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 primary purpose of adapting LMS labels for this ordinal categorical scale was to leverage the well-established verbal descriptors to help assessors interpret and apply the intensity categories consistently, rather than to replicate the quasi-logarithmic perceptual spacing of LMS. Including 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 labelled 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).

This approach prioritised ease of practical usability over preserving the mathematical properties of the original LMS. It is important to note that while the category labels are evenly spaced, the Rasch modelling approach does not assume these categories represent equal perceptual intervals. Rather, the model empirically estimates the threshold parameters between each category based on actual response patterns in the data, transforming the ordinal ratings into interval-level logit measures (Bond et al., 2020; Eckes, 2023). This means that the perceptual spacing between categories is calibrated based on how the panel actually used the scale to rate the attributes across the samples, rather than imposing uniform intervals.

Assessors were provided with a cup of water to cleanse their palate between sample evaluations and given breaks between replicates (5 minutes for the DFC and 10 minutes for the AR test) to minimise sensory fatigue and memory bias, respectively. Samples of the questionnaires for the DFC and AR tests are provided in Appendix **C.1**.

---

[3] *Step three* (in Figure 3.1:Framework): choose an intensity rating scale to categorise attribute intensities into ordinal scores, and

[4] *Step four*, collect ratings through sensory evaluation.

4.2.5 Data analysis

Rasch and statistical analyses were according to the procedures described in the previous chapter (in section **3.3**). [5]The attribute intensity ratings (AR) data were fitted to a MFRM with four facets: *Assessors*, *Samples*, *Repetition,* and *Attributes*. To facilitate the comparison between the two approaches, a separate model was used for the DFC data. Each test had two variations of the model, one with and one without the *Repetition facet* as outlined below. **TIM1** and **DFCM1** models include all four facets, with data generated from all three replicate sessions grouped under the *Repetition facet*, while **TIM2** and **DFCM2** models exclude the *Repetition facet* and instead fit the data from individual replicated sessions of both tests to the MFRM. This was necessary to investigate whether assessors provide consistent ratings within single sessions or if averaging across multiple replicate sessions is required for reliable measurement.

TIM1: $\boldsymbol{ln\,(P_{mnrik}/P_{mnrik-1}) = \beta_m - \theta_n - \rho_r - \delta_i - \tau_k}$

…Equation 4.1

TIM2: $\boldsymbol{ln\,(P_{mnik}/P_{mnik-1}) = \beta_m - \theta_n - \delta_i - \tau_k}$

…Equation 4.2

DFCM1: $\boldsymbol{ln\,(P_{mnrk}/P_{mnrk-1}) = \beta_m - \theta_n - \rho_r - \tau_k}$

…Equation 4.3

DFCM2: $\boldsymbol{ln\,(P_{mnk}/P_{mnk-1}) = \beta_m - \theta_n - \tau_k}$

…Equation 4.4

Where: in the DFC models (DFCM), the δ_i parameter was not included due to the absence of attributes in the analysis.

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

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

β_m= degree of leniency or severity of assessor (m) in rating attribute intensities

θ_n= degree of difference in the total intensity measure for sample (n)

ρ_r = degree of difference between ratings of samples in a replicate session (r)

[5] *Step five* (in Figure 3.1:Framework): fit a MFRM to obtain interval-scaled Total Intensity Measures (TIM) based on combined attributes that will be used for univariate statistical analysis.

δ_i = the average degree of intensity of sensory attribute (i) across all samples

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

Statistical analyses were conducted on the DFC raw scores, DFC Rasch measures, and the Total Intensity Measures (TIM), and the results were compared for discriminatory ability and diagnostic detail.

4.2.5.1 Rasch Model Fit

To recap, an acceptable global model fit of the data is when no more than 5% of absolute standardised residuals is ≥2, and no more than 1% is ≥3 (Linacre, 2022; Eckes, 2023).

For individual fit of each parameter within each facet (i.e., *Assessor, Sample, Repetition* and *Attributes*) adequate model fit is assessed using OUTFIT mean square values. Values between 0.5 and 1.5 are considered useful for measurement, while values >2.0 may degrade the measures, and values <0.5 may indicate redundancy or insufficient discrimination (Linacre, 2025b).

"Response dependency" checks (Tennant & Conaghan, 2007) were conducted to examine unidimensionality (i.e., ensuring that attributes are measuring a single construct) and local item dependence (i.e., ensuring that responses to different attributes are not overly correlated unless they are truly measuring the same thing, making them redundant). Linacre (2024a) suggests using a Principal Component Analysis of Residuals (PCAR), where unidimensionality is confirmed when the eigenvalue of the unexplained variance in the first contrast is <2. Local item dependence is identified when the residual correlation between two attributes is >0.3 (Ramp et al., 2009; Christensen et al., 2017). Response dependency checks were only applied to the **TIM1** model, as it is based on multi-attribute responses and is the proposed model for difference testing.

4.3 Results and Discussion

4.3.1 Fit of data to the Many-Facet Rasch Model (MFRM)

Rasch model fit statistics were examined for all fitted models (TIM1, DFCM1, TIM2 and DFCM2) to determine whether the data support unidimensional measurement.

Table 4.1. 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 (subsequently OUTFIT Mnsq) indicating deviation of the estimates from predictions of the Rasch model.

[2] Percentage (number of observations in brackets) of absolute standardised residuals (StRes).

[3] Total number of observations used for the estimation of the model parameters.

[4] OUTFIT Mnsq values between 0.5 and 1.5 are considered productive for measurement (Linacre, 2024b). The same criteria apply to the percentage fit for all facets.

[5] OUTFIT Mnsq values >2.0 may degrade the measurement (Linacre, 2024b).

[6] NA implies Not Applicable as the Rasch models per replicate did not have a Repetition facet.

As shown in **Table 4.1** and **Table 4.2**, all models showed an acceptable global model fit suggesting that overall, the data in each model aligns with the assumptions of the Rasch model and there are no major inconsistencies that may distort the measurement.

Evidently, all facets across all fitted models, except the *Assessor facet*, showed a 100% fit. Assessor fit indices estimate how consistently an assessor's ratings align with the expectations of the model. (Myford & Wolfe, 2004; Linacre, 2012a; Eckes, 2023). While there is evidence of a few misfitting assessors across all models, the focus of this chapter was on investigating MFRM's ability to measure sample differences rather than comprehensive assessor performance evaluation. Monitoring assessor performance with the MFRM is discussed in **Chapter 5**.

Moreover, according to (Wright & Linacre, 1994), a few misfitting assessors in sample and item (attribute) estimates are negligible. Notably, the TIM1 and DFCM1 models with the *Repetition facet* showed better assessor fit than the corresponding models without the *Repetition facet*. This suggests that including *Repetition* as an explanatory factor and averaging across replicated ratings helps smooth out random variations, making the data from these models more reliable. Meilgaard et al. (2025) highlight repeating measurements as one of the techniques to minimise variability in product ratings due to individual differences.

Response dependency checks (**Table 4.2**) on the TIM1 model confirmed that the combined attributes formed a unidimensional measurement construct. PCAR showed that after removing the Rasch factor*, the unexplained variance in the first contrast (representing residuals in the largest secondary dimension) with an eigenvalue of 1.9 indicated a strength of 2 out of 5 items, suggesting the possibility of a secondary dimension. However, examination of the standardised residuals correlation matrix showed that correlations between suspected attributes were <0.3 indicating that any observed associations were weak and likely due to local variations in attribute intensity (Linacre, 2024a). That said, the observed association between *Orange flavour* and *Sweetness* could theoretically reflect sensory interactions where citric acid can enhance sweetness perception (Veldhuizen et al.,

* The Rasch factor is the primary dimension representing the latent trait measured by the Rasch model and reflects the expected response pattern.

2017). However, the correlation value of 0.05 is negligible and more likely reflects random measurement variation. Meanwhile, the weak correlation between *Milky flavour* and *Saltiness* (0.11) may be attributed to their similarly low intensities and erratic ratings by assessors, as discussed later in the chapter.

Table 4.2. Summary of response dependency based on standardised residuals

Model	**Response Dependency**		
	Unidimensionality[1]	**Local Item Dependence**	
		Attributes	**Corr. of StRes**[2]
TIM1	1.9	Milky flavour-Saltiness	0.11
		Orange flavour-Sweetness	0.05
Criteria	*eigenvalue <2 in 1st contrast*		*<0.3*

[1] Eigenvalue of the unexplained variance in the first contrast, not accounted for by the Rasch model, in the Principal Component Analysis of Residuals (PCAR).

[2] Correlation of standardised residuals (Corr. of StRes) <0.3 confirm responses on attributes are not related.

4.3.2 Rating scale category diagnostics

Scale category diagnostics, one of the many quality control statistics offered by Rasch models provide insights into how the categories on a rating scale have been interpreted. Following established guidelines in **Table 3.1: Guidelines for assessing the functionality of a rating scale**, deviations in the interpretation and operational use of the scale from the Rasch model's expectations can be empirically investigated.

Table 4.3 shows the category functioning of the rating scales for the Intensity and DFC rating scales for the Rasch models that include a *Repetition facet*- TIM1 and DFCM1 respectively. The scales were examined against the previously discussed criteria (**Table 3.1**). Criteria which were essential for measure accuracy and for description of the samples in this study were met.

Table 4.3. Summary of scale category statistics for Intensity and DFC rating scales used in the TIM1 and DFCM1 models (with *Repetition facet*)

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 Mnsq 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 (i.e., the thresholds between adjacent scale categories. For 8 and 7 category scales, the minimum threshold distances are 0.51 and 0.57, respectively.

Unmet Criteria from 3.3.1.4: Rating scale category diagnostics

* Each scale category should have at least 10 observations as this is essential for measure stability.

^ Minimum advancing distance for Rasch-Andrich threshold are helpful for inference on subsequent studies

Specifically, the Rasch-Andrich thresholds were ordered, and probability curves had distinct peaks (see **Figure 4.3** for graphical representations of the probability curves for the Intensity scale). No misfitting categories were observed, as OUTFIT Mnsq values were close to 1.0, and the observed average measures increased monotonically across the scale categories. These findings suggest that responses to attributes in the TIM are consistent with the estimates of the latent variable (Tennant & Conaghan, 2007) and meet the model expectations. Additionally, no scale categories were skipped along the variable (Eckes, 2023).

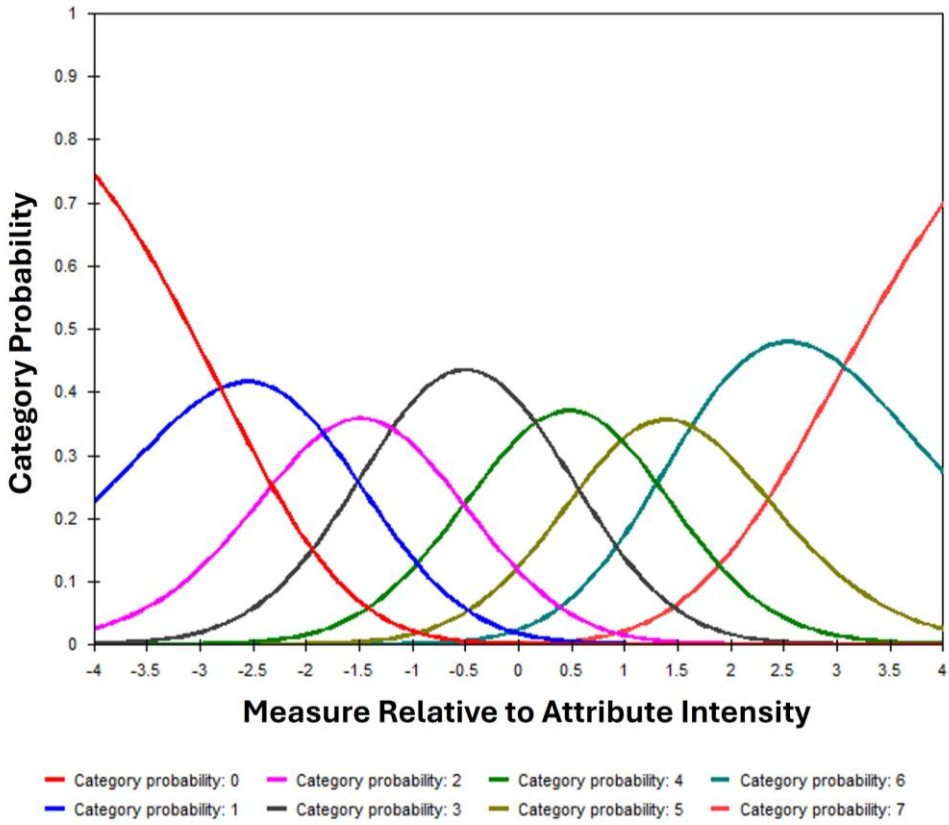


Figure 4.3. Probability curves for TIM1 Intensity scale showing ordered Rasch-Andrich thresholds resembling a range of hills with distinct peaks. As scale categories advance along the latent variable, each category becomes the most probable choice. The points where each category curve intersects with the adjacent category curve represent the half-point (Measure in **Table 4.3**), where the probability of a sample receiving a higher rating begins to exceed the likelihood of being rated in the lower adjacent category. These correspond to the half-point thresholds on the Wright maps.

However, there were only 6 total observations in the last category (7 - Strongest imaginable oral sensation) of the Intensity scale. According to Linacre (2002b), a minimum of 10 observations per category is essential for ensuring measure

stability, which refers to the consistency of a measurement system when repeated over time in the same context.

The DFC scale also did not meet the required minimum advancing distance between category thresholds for scales with 7 categories. Specifically, the thresholds categories 2 - Slight difference and 5 - Very large difference, respectively were less than the minimum required 0.57, suggesting that these categories were too close to be distinctive (Eckes, 2023). However, meeting this requirement is only helpful for making inferences in subsequent studies. Therefore, it was not necessary to revise either of the rating scales, as doing so would have been beyond the scope of this study, which focused primarily on exploring the MFRM for measuring overall differences, rather than modifying tools to improve measurement procedures for Jaffa cakes.

4.3.3 Representing the Overall Difference construct on the Wright map[6]

Wright maps for the TIM1 and DFCM1 models are presented in **Figure 4.4** and **Figure 4.5** respectively. As previously discussed in section **3.3.1.1: Fitting the Many-Facet Rasch Model (MFRM)**, all four facets (*Assessors*, *Samples*, *Repetition* and *Attributes*) were positively oriented so that on average, for each facet the following applies.

- **Assessor facet:** assessors with higher logit values are more lenient, generally assigning higher scores on the rating scale.
- **Sample facet:** samples with higher logit values have higher Total Intensity Measure (TIM) or for the DFC measure (DFCM), are more different from the control.
- **Repetition facet:** replicate sessions where higher intensity ratings were assigned on average have higher logit values.
- **Attribute facet:** attributes with higher average intensity ratings have higher logit values.

[6] *Step six* (in Figure 3.1:Framework): represent the construct on the Wright map to visualise the location of facet parameters on the logit scale. Rasch measures from *steps five* and *six* are exported for statistical analysis, and Rach quality control statistics (OUTFIT Mnsq) provide insights into specific attribute contribution to the latent variable of overall difference.

The individual parameters within each facet (e.g., each assessor in the *Assessor facet*, each sample in the *Sample facet*, etc) are relatively located on the Wright map according to their logit values. The *Sample facet* was non-centred, while the other facets were centred at the mean (0 on the logit scale) to serve as a reference point. Consequently, sample locations were adjusted by considering the severity of assessors, the average intensity of attributes, and the intensity ratings in repeated sessions representing the *Assessor, Attribute*, and *Repetition facets*, respectively.

4.3.3.1 Total Intensity Measure (TIM1)

The TIM Wright map (**Figure 4.4** above) showed that assessors exhibited varying degrees of severity in their use of the intensity rating scale. On the *Assessor facet*, Assessor 1014 had the highest logit value and emerged as the most lenient assessor in the panel. This suggested that they consistently assigned the highest ratings to the samples compared to other assessors.

For the *Sample facet*, 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 positioned higher on the scale were perceived to have greater intensity of the combined attributes. Their values on the logit scale, relative to their location, represent the Total Intensity Measure, which will be used for multiple comparison tests and reflects the latent variable of Overall Difference between the samples.

The *Attributes facet* revealed the location of the attributes based on the average ratings from the samples. Attributes higher on the scale had the highest intensity ratings across all the samples and repeated sessions. *Orange flavour* and *Sweetness* had the highest average intensity, while Milky flavour and Saltiness had the lowest.

Initial interpretation suggested this hierarchy indicated how much each attribute contributed to differences between samples, with *Orange flavour* and *Sweetness* appearing most influential. However, further examination using MFRM across different contexts revealed that this hierarchy actually reflects average intensity levels rather than discriminating power. Attributes with high average intensity are more easily perceived overall and do not necessarily mean they contribute most to sample differentiation.

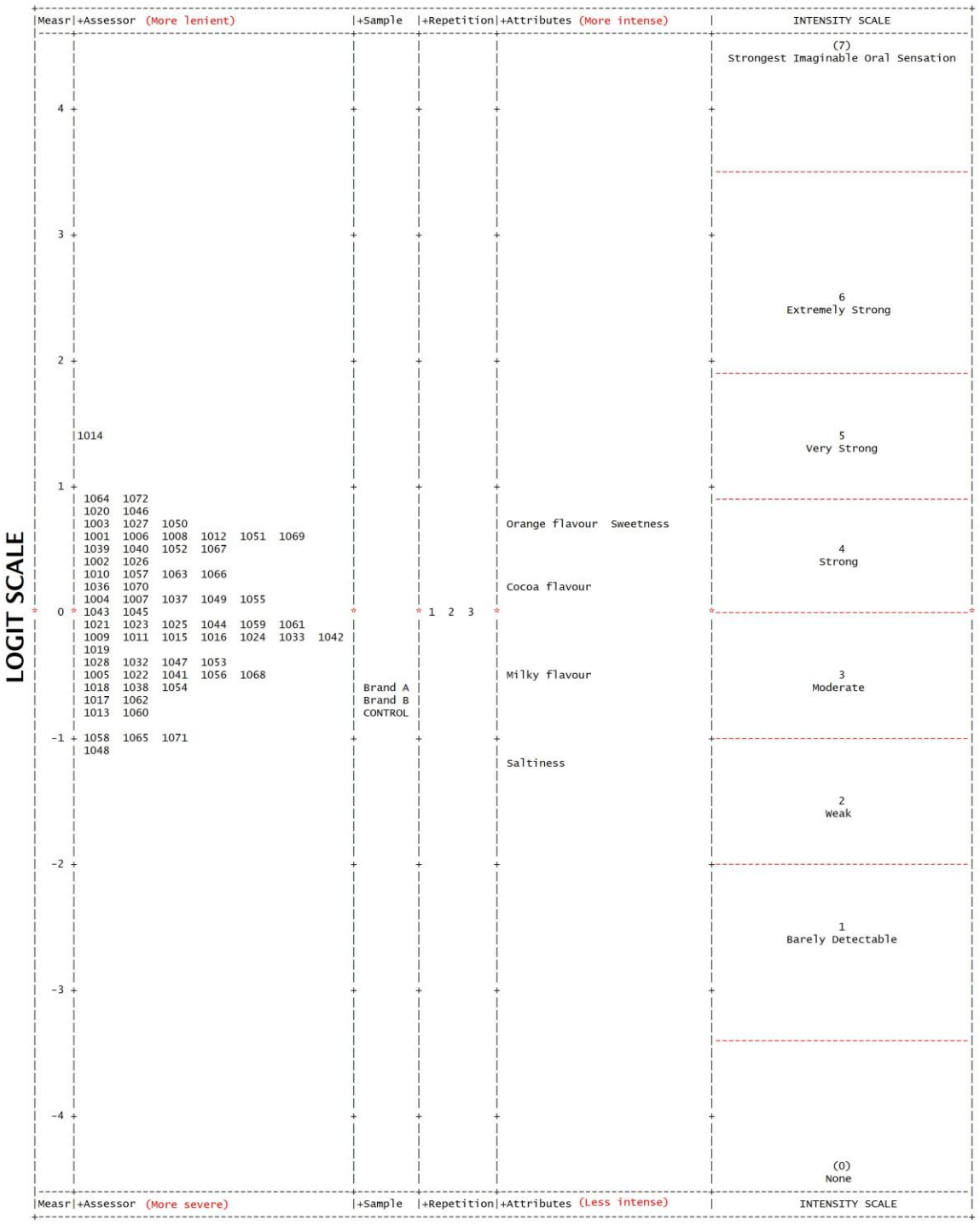


Figure 4.4. Many-Facet Wright map for TIM1.

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.

The interpretation of attribute intensities in the MFRM depends on the measurement context and the construct being modelled. In Ho (2019), the construct was overall liking, where individual attribute intensities served as items contributing to that liking judgment. In that context, attribute locations on the logit scale directly indicated their contribution to overall liking, as higher attribute intensities translated to higher liking scores. In contrast, the construct examined in this thesis is overall difference between samples, with attribute intensities also serving as items. Here, higher logit values reflect higher average intensity across all samples rather than a stronger contribution to sample discrimination. When statistically significant differences exist between samples, the Outfit mean square values for the attributes (discussed later in this chapter) provide insight into their relative contributions to those differences.

The intensity scale shows the average rating range used by the panel for the attributes. Notably, the gaps between adjacent scale categories are not equidistant, and tend to widen toward the extreme categories. The Rasch model empirically estimates these category thresholds from the observed response patterns as rated by the assessor panel rather than imposing uniform spacing, with the non-uniformity reflecting how assessors actually used the scale categories when rating the samples. On average, all samples were rated as having moderate intensity across the combined attributes.

To estimate the overall difference analogous to the DFC method, 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. Rasch quality control statistics, specifically the OUTFIT mean square for individual attributes, would further reveal the importance of each attribute to the overall sample differences.

4.3.3.2 DFC Measure (DFCM1)

In **Figure 4.5** below, the DFCM1 Wright map revealed varying degrees of severity among assessors in the *Assessor facet*. Assessor 1011 was the most severe, consistently assigning the lowest ratings to samples on average, standing out from the rest of the panel by nearly 2 logits. This extreme severity indicates they used the

rating scale very differently from their peers, systematically rating all samples lower. Rasch quality control statistics would flag this assessor as misfitting due to their deviation from the model's expected response patterns.

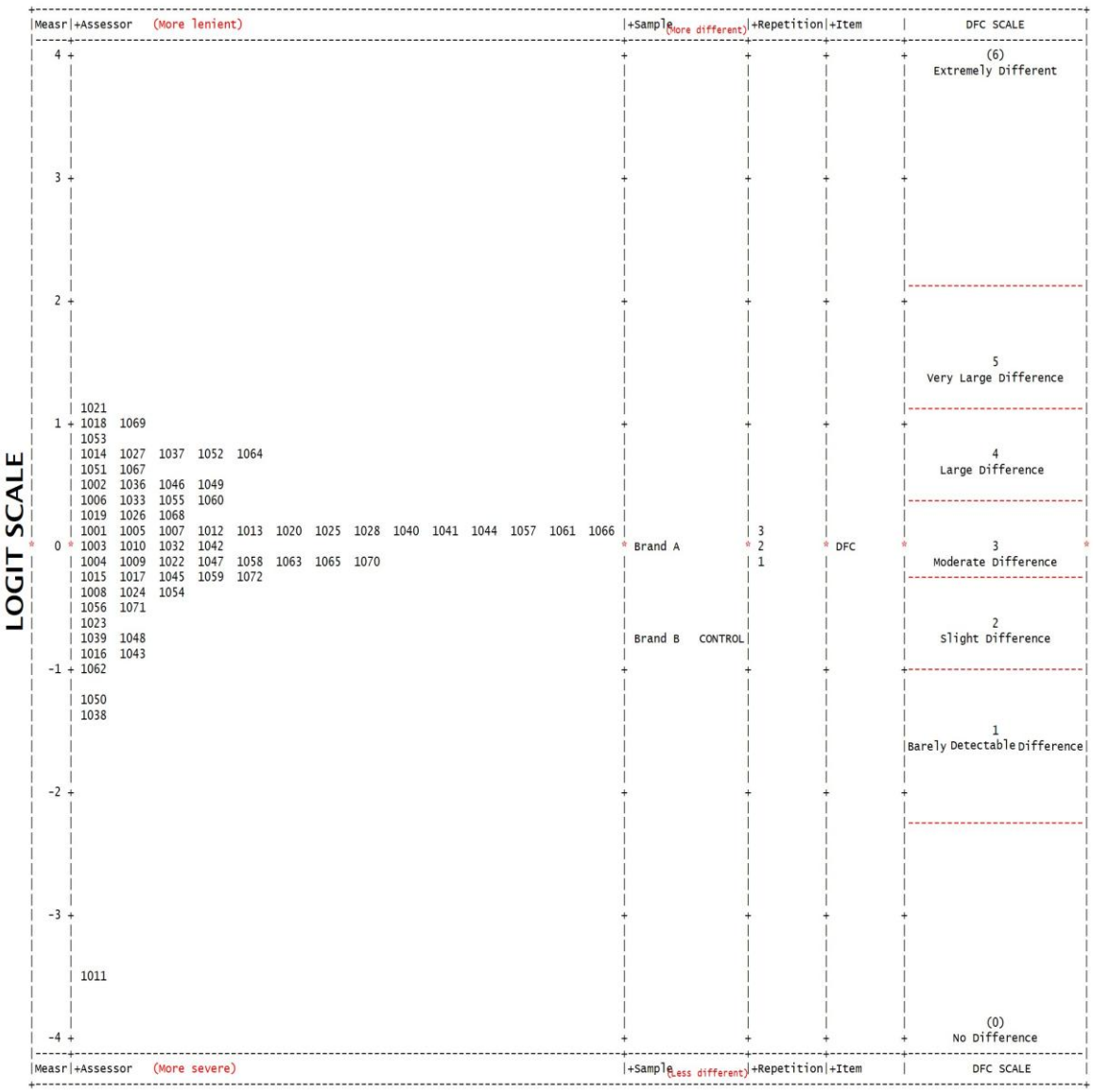


Figure 4.5. Many-Facet Wright map for DFCM1.

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. Numbers 1-3 indicate replicate evaluations in the repetition facet, 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.

This type of assessor behaviour, where an individual systematically rates differently from the panel consensus, poses challenges for traditional sensory analysis methods. In ANOVA approaches, such assessors contribute to increased error variance, but their

specific problematic patterns may not be readily identified without additional tests. MFRM, in contrast, provides immediate individual-level fit statistics that highlight these issues, enabling targeted assessor retraining or data quality decisions. A detailed evaluation of assessor performance using MFRM, including the implications of different types of misfit patterns, is addressed in **Chapter 5**.

For the *Sample facet*, Brand A was notably located higher on the logit scale compared to Brand B and the CONTROL which had similar logit values. While on average, assessors rated Brand A as moderately different from the CONTROL, the slight difference rating between Brand B and CONTROL was not statistically significant. Some assessors may have considered differences in other sensory attributes across different modalities such as appearance, texture, and other flavours, which were not intended to be captured in the study. Efforts to maintain consistency across other sensory characteristics, aside from those of interest, during sample selection may not have been entirely successful. Brand A had a slight difference in shape compared to the other samples (**Figure B 1**), which some assessors may have noticed. Furthermore, post-study feedback revealed that several participants were able to 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 due to assessors probably experiencing fatigue and some context bias from tasting numerous samples during the test. As fatigue and cognitive overload set in, assessors may simplify their responses by restricting the range of their ratings to a particular section of the scale. Due to sensory adaptation, this restricted range may shift toward the higher end of the scale. As noted by Lawless and Heymann (2010) and Meilgaard et al. (2025), repetitive and demanding testing conditions can compromise panel performance, leading to increased response variability and a reliance on habitual rating patterns.

As with the TIM Wright map, the gaps between adjacent scale categories are not equidistant and tend to widen toward the extreme categories. As reported by (Tennant & Conaghan, 2007), values at the extremes of the scale capture a wider range of the underlying construct, in this case, the difference of the samples from the control (DFC).

4.3.4 Comparing overall difference between samples: DFC versus TIM

The overall difference results from DFC raw scores were compared with those from the Rasch-based TIM as well as to DFC Rasch measures (DFCM) for their sensitivity to product differences and the level of diagnostics information they provide.

Table 4.4 below summarises the statistical test results for DFC raw scores, TIM and DFCM, together with their replicates. 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 (Wright & Masters, 2002; Myford & Wolfe, 2003). 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, 2002).

All datasets for the DFC Rasch measures showed Strata values greater than 4, indicating that the model could reliably distinguish multiple statistically distinct levels of perceived DFC among the samples. Reliability values close to 1.0 further support the precision of the measures. These indicators reflect strong overall discriminatory ability, as supported by the Wright map, which showed that Brand A was notably different from the Control and Brand B, being located approximately 0.5 logits away from them. Pairwise comparisons would help identify which specific samples differed significantly from the control.

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 of 0.45 and 0.35, respectively. These low values suggest that, in the first and second evaluations of sample replicates, assessors were unable to reliably distinguish between the samples.

Table 4.4. Comparison of Sample facet summary statistics for all TIM and DFCM Rasch models and raw DFC scores, with mean comparisons based on the Friedman test.

	TIM Models				DFCM Models				DFC RAW Scores			
Test/Dataset[1,2]	TIM2.R1	TIM2.R2	TIM2.R3	**TIM1**	DFCM2.R1	DFCM2.R2	DFCM2.R3	**DFCM1**	**R1**	**R2**	**R3**	**R. Avge**
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**	5.4	5.01	7.86	**8.78**				
Rasch Fixed χ^2_{Sample}	**5.4$^{\sim}$**	**4.6$^{\sim}$**	**9.4****	**17.9*****	**48.1*****	**41.0*****	**101.9*****	**128.6*****				
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	**0.29**
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**	0.07	<0.001	<0.001	**0.271**	<0.001	0.011	0.002	**<0.001**
Friedman Test												
χ^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***	0.2	**0.05**	0.01	-0.01	0.18	**0.06**

[1] P-value levels of significance: <0.001***, <0.01**, <0.05*, <0.1$^{\sim}$; measures with no superscript symbols have p-values >0.1.

[2] NA indicates "Not applicable" as no outliers were found.

TIM = Total Intensity Measure; DFCM = Difference-from-Control Model. TIM1 and DFCM1 = models including the Repetition facet; TIM2 and DFCM2 = models excluding the Repetition facet (i.e. individual replicate datasets). R1–R3 = Repetitions 1–3 for the corresponding model; Raw DFC scores are presented for individual replicates (R1–R3) and for the averaged values. R. Avge = average across repetitions.

The Reliability values below 0.5 indicate the dominance of measurement error (Wright & Masters, 2002; Myford & Wolfe, 2003; Linacre, 2023b), suggesting more inconsistent ratings during the earlier replicate sessions, possibly due to initial uncertainty or unfamiliarity with the methods and samples.

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, the TIM1 model which included the fourth facet for repetition by combining the three repeated sessions, showed a Strata value of 3 indicating three statistically distinct levels among the samples, supported by a Reliability value closer to 1.0. This suggests that modelling all replicate sessions simultaneously and accounting for variability from repeated evaluations reduced inconsistencies in assessor ratings and thereby improved the discriminatory ability of the measurement. The model accounted for the variability by estimating separate parameters for the repetition effect or facet, effectively removing its influence from the sample comparisons.

The differences in separation statistics between the DFCM and TIM models likely reflect differences in the constructs being measured. DFCM captures a holistic judgement with a single item (overall Difference From Control), whereas TIM assesses five separate attribute intensities. As a result, the observed separation values may reflect genuine differences in how discriminable the samples are along these distinct measurement dimensions.

Test design and cognitive strategy also likely influenced discrimination. The DFCM used a comparative presentation in which assessors directly evaluated each sample against a control, a process that can support more consistent responses. In contrast, TIM used monadic presentation, requiring assessors to rate five attributes separately without the control present and to rely on their own internal reference for each attribute. The overall difference score for TIM was then derived as a latent variable from these individual ratings.

The higher Strata values observed for DFCM may additionally reflect the substantial perceptual differences between the products assessed, particularly the strong contrast between Brand A and the control.

These approaches serve different research purposes. DFC was chosen as the comparison because it is the established method for directly measuring overall difference and quantifying its magnitude. It produced higher separation in this study, making it efficient for determining whether products differ. In contrast, TIM derives overall difference from individual attribute ratings and therefore provides diagnostic information about which specific attributes contribute to the perceived differences. This makes it particularly useful when the research goal extends beyond simple differentiation to understanding attribute contributions to product differences. The trade-off is that TIM's predefined attributes may not capture all perceptible differences detected in holistic DFC judgments. Therefore, when using TIM, test design must ensure that relevant attributes are carefully selected and included in the sensory questionnaire to allow accurate estimation.

Parametric two-way ANOVA tests also indicated statistically significant differences between samples across all datasets ($p<0.001$), except for TIM2.R1 which was only marginally significant ($p<0.10$), and TIM2.R2 with a slightly greater significance ($p<0.05$). These results are consistent with the Rasch model fixed chi-square statistics and separation indices, which also showed weaker model performance for these replicates, with strata values below 2 and reliability values under 0.5, as previously discussed. However, residual analyses revealed violations of key ANOVA assumptions **(Table 4.4)**. Specifically, non-normality was detected in both the TIM and DFCM estimates, and Breusch-Pagan tests confirmed heteroscedasticity in residuals across all datasets (DFC raw scores, DFCM, and TIM models). As a result, non-parametric methods were employed.

In earlier analyses, as reported in (Ariakpomu et al., 2025b), differences between samples were assessed using non-parametric mean comparison and post hoc tests that included sample x assessor interaction effects. While this approach is commonly used to detect subtle differences in raw score data, it is not suitable for Rasch-derived measures such as TIM and DFCM, unless the Rasch model used explicitly includes the assessor x sample interaction effects. This is because the Rasch estimation process already adjusts for each assessor's severity or leniency when generating measures. Rasch model estimates account for assessor effects, as well as those of other modelled facets, effectively removing their influence.

Reintroducing these assessor effects in rank-based mean comparison tests, such as the Friedman test, leads to double-counting the variance. This makes the test overly sensitive, increasing the risk of *Type I errors*, as the between-group variations have already been separated out during the Rasch modelling process. Any remaining unexplained variations, including interaction effects are modelled as measurement error and reflected in the residuals (Linacre, 1995).

This issue is evident in **Table 4.4**, where applying the Friedman test produced a chi-square value of 134, and indicated a highly statistically significant difference ($p<0.001$) between the samples across the Rasch-derived measures (TIM and DFCM models). Any observed statistical significance may be reflecting redundant variance rather than true differences between samples. This is because the Friedman test treats assessors as blocks, ranking the samples within each assessor based on the raw scores they provided, and then compares these ranks across assessors to test for differences. However, when the Rasch model already reveals significant differences in assessor severity levels, as with this study, those effects have already been statistically adjusted for in the estimation process. Specifically, the model includes rater severity as a separate parameter, so systematic variance associated with individual assessors is modelled and removed from the resulting Rasch measures. Residual assessor inconsistency not captured by the severity parameter is absorbed as measurement error in the Rasch residuals. Applying the Friedman test to these adjusted measures therefore reintroduces assessor effects that the Rasch model has already accounted for, exaggerating the detection of between-sample differences.

To illustrate, consider an excerpt from the Rasch-derived Total Intensity Measures (TIM) for the three Jaffa cakes samples, based on evaluations from three assessors:

Assessors/Samples	Brand A	Brand B	Control
Assessor 1	0.05	-0.15	-0.22
Assessor 2	-0.16	-0.27	-0.35
Assessor 3	0.12	0.00	-0.07
Assessors ... n=67(mean)	**-0.59**	**-0.71**	**-0.78**

While the absolute values differ between assessors due to differences in severity, the within-assessor ranking of the samples remains consistent: Brand A > Brand B

> Control. Because the Friedman test operates on these within-assessor block ranks, it will consistently indicate significant differences between the sample ranks. Analyses of additional datasets further corroborated this issue (see **Table E 1**), consistently showing inflated significant sample differences with the Friedman test. Since Rasch-derived measures already adjust for rater severity and other modelled sources of variance (Boone et al., 2014; Bond et al., 2020), post hoc comparisons for Rasch measures should focus solely on differences between samples and avoid including assessors as blocking factors. Where parametric assumptions are unmet, non-parametric alternatives such as the Kruskal-Wallis and Dunn's tests are more appropriate.

Using these revised methods, the findings presented in **Table 4.5** more accurately reflect the sample differences and align with the results of the Rasch fixed chi-square statistics. As noted by Boone et al. (2014), when data fit the Rasch model, the resulting interval-level measures are generally suitable for parametric analyses. However, because sensory data sometimes violate parametric assumptions (Kemp et al., 2018), non-parametric alternatives can still be used with confidence when applied to Rasch measures, not as a compromise, but as equally robust options. This robustness stems from the model's use of raw scores as "sufficient statistics" (Linacre, 2004b; Bond et al., 2020), meaning that the total score contains all the necessary information to estimate the location of the person or items (in this context, the samples or attributes) on the latent trait, based solely on the structure of response patterns rather than assumptions about the underlying distribution.

The sample differences identified by the Kruskal-Wallis test, which evaluates group differences based on differences in mean rank sums per sample, closely aligned with the results from the Rasch analysis (Rasch fixed χ^2). All results, except the individual replicates of the TIM model (TIM2), indicated that the difference between Brand A and the Control was highly significant ($p < 0.001$). For the TIM2 model, the degree of significance increased progressively across replicates 1 to 3, with corresponding Dunn's test p-values of 0.14, 0.08, and 0.003, respectively. This pattern likely reflects a learning effect where repeated exposure reduced initial uncertainty and allowed assessors to develop a stable internal frame of reference, enhancing their ability to discriminate subtle sensory differences.

Table 4.5. Comparison of Sample facet summary statistics for all TIM and DFCM Rasch models and raw DFC scores, with mean comparisons based on the Kruskal-Wallis test.

	TIM Models				**DFCM Models**				**DFC RAW Scores**			
Test/Dataset[1]	TIM2.R1	TIM2.R2	TIM2.R3	**TIM1**	DFCM2.R1	DFCM2.R2	DFCM2.R3	**DFCM1**	**R1**	**R2**	**R3**	**R. Avge**
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**	5.4	5.01	7.86	**8.78**				
Rasch Fixed χ^2 $_{Sample}$	**5.4$^{\sim}$**	**4.6$^{\sim}$**	**9.4****	**17.9*****	**48.1*****	**41.0*****	**101.9*****	**128.6*****				
Mean Comparison Tests												
Friedman Test (χ^2)[2]	134***	134***	134***	134***	134***	134***	134***	134***	20.39***	14.21***	45.80***	46.72***
Kruskal-Wallis Test (H)[3]	**3.94**	**5.09$^{\sim}$**	**11.78****	**18.99*****	**38.33*****	**42.21*****	**78.91*****	**210*****	**20.50*****	**18.18*****	**44.52*****	**79.12*****
Mean differences (Dunn's Many-to-One Test)												
Control-Brand A	-0.19$^{\sim}$	-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$^{\sim}$**	0.02	-0.01	0.2	**0.05**	0.01	-0.01	0.18	**0.06**

TIM = Total Intensity Measure; DFCM = Difference-from-Control Model. TIM1 and DFCM1 = models including the Repetition facet; TIM2 and DFCM2 = models excluding the Repetition facet (i.e. individual replicate datasets). R1–R3 = Repetitions 1–3 for the corresponding model; Raw DFC scores are presented for individual replicates (R1–R3) and for the averaged values. R. Avge = average across repetitions.

[1] P-value levels of significance: <0.001***, <0.01**, <0.05*, <0.1$^{\sim}$; measures with no superscript symbols have p-values >0.1.
For degrees of freedom (df) = 2, the chi square (χ^2) critical values are 5.991 ($\alpha = 0.05$) and 4.605 ($\alpha = 0.1$).

[2] Friedman test results are included for comparison only. The inflated significance reflects redundant variance already modelled by the Rasch estimation.

[3] The Kruskal-Wallis test statistic (H) provides the primary analysis for between-sample differences and also follows a χ^2 distribution for significance testing.

This effect is consistent with findings by Peltier et al. (2018), who observed that replicating evaluations enhances the ability to discriminate between flavour attributes. In contrast, the DFCM results showed high discrimination from the first evaluation, likely because assessors directly compared each sample against a physical reference (control) rather than relying solely on a mental internal reference. The comparative design of the DFC test inherently provided a physical reference that facilitated immediate discrimination, leading to higher Strata and reliability estimates, whereas the monadic presentation of the attribute rating test used for TIM required repeated exposures for assessors to construct an equivalent internal reference, explaining the progressive increase in sensitivity across replicates.

Beyond presentation design, the higher Strata values for DFCM compared to TIM may also reflect that the holistic DFC rating allowed assessors to integrate any perceptible difference into their judgment. When rating overall difference from the control with a physical reference present, assessors could detect differences across multiple sensory modalities, not just taste. However, these ratings could have been influenced by perceived differences other than the taste of the samples. As previously discussed in **4.3.3.2: DFC Measure (DFCM1)**, the perceived difference in non-taste attributes and familiarity with Brand A may have influenced assessors' DFC ratings, despite efforts to minimise these influences. In 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 help distinguish the samples more effectively based on combined ratings, as will be explored in **Chapter 6**.

4.3.5 Examining attribute contributions to the overall difference (TIM)

This section examines how individual attributes contribute to the overall difference as a latent variable using the TIM model. Rasch outfit mean squares for the *Attribute facet* were used to assess how well each attribute distinguished between the samples. As discussed earlier in **Chapter 3**, Rasch residual fit statistics - Infit and Outfit evaluate how well data associated with individual parameters in a facet align with the expectations of the Rasch model. To recap, as a general rule (Linacre,

2024b), after accounting for measurement error, mean square values less than 1 (<1) indicate *overfit,* meaning the observed ratings are more predictable than the model expects. Conversely, values greater than 1 (>1) indicate *underfit,* where observed values deviate more from the model's expectations. However, fit criteria have been found to be context dependent (Wu & Adams, 2013; Eckes, 2023) and so acceptable ranges should be set accordingly as detailed in **3.3.1.6:Residual fit statistics.**

For the Item facet, the Rasch model assumes equal discrimination across all items. The OUTFIT statistic measures how well each item's response patterns fit the model's expectations, i.e. whether an item's response pattern deviates from the expected pattern of equal discrimination across items (Wu & Adams, 2013), indicating how much unexpected variation there is in the response data for item. In this study, "items" refer to the sensory attributes, and the Outfit mean square indicates how much variation is present in the ratings assigned to an attribute by assessors across the different samples. Therefore, variations in attribute ratings between samples may occur due to one or more of the following: actual perceptible differences between samples, individual differences in assessor perception, or unclear or inconsistently understood attribute definitions. The results provide a high-level indication of where variations in attribute ratings occur across the samples, guiding further investigation into potential underlying causes, and whether these variations are primarily due to differences between the samples or assessor bias.

The OUTFIT mean square for each attribute can indicate the following.

- **Values below the acceptable range (overfit):** The attribute may not discriminate well between samples, as the responses are overly predictable.
- **Values above the acceptable range (underfit) in high-intensity attributes:** This suggests that the discrimination of the attribute differs significantly from that of other attributes (Wu & Adams, 2013). According to (Linacre, 2025b), when an easy item (i.e. a high-intensity attribute) shows underfit, it may indicate that the item behaves qualitatively different from the others.

- **Underfit in low-intensity attributes:** When low-intensity attributes (i.e., those with negative logit values or values below the mean) have Outfit mean square values above the acceptable range (underfit), in most cases it may reflect inconsistency in how assessors rate the attribute. Inconsistent ratings could result from individual perceptual differences, ambiguous attribute definitions, or response bias, where assessors interpret the attribute in different ways or are uncertain whether they actually perceive the attribute. This uncertainty may lead to considerable variation in ratings across repeated sessions (i.e., low internal consistency). Linacre (2025b) reports that when a difficult item (in this context, a low-intensity attribute) is underfit, it is often ambiguous, debatable or contains misleading options.

  However, this is not always the case. Underfit may still reflect genuine variation in the attribute across samples, particularly when an attribute receives consistently low ratings due to its absence in one or more samples. This can distort the overall ratings, pulling the attribute to the low end of the logit scale, as the low rating for a sample consequently lowers the average ratings across all samples. Careful interpretation is therefore required to distinguish whether underfit is due to assessor inconsistencies, rating distribution issues or inherent attribute characteristics.

When differences between samples are statistically significant, higher Outfit mean square (Outfit Mnsq) values suggest that the attribute was contributing more to the observed differences between samples. In contrast, Outfit Mnsq values below the acceptable range indicate that the attribute shows little variation across samples, possibly because it was redundant or measured with limited sensitivity. This lack of variability suggests that the attribute does not effectively differentiate between the samples after accounting for measurement error.

In this study, the total number of responses (Nr) for each attribute was 603, hence the acceptable OUTFIT Mnsq range was calculated* (as discussed in **3.3.1.6**:**Residual fit statistics**) to fall within 0.89 -1.12. **Table 4.6** presents the OUTFIT Mnsq results from the Rasch analysis, alongside results from the three-way

* $1 \pm 2\sqrt{\frac{2}{Nr}}$, where Nr (number of responses) for each of the attributes is 603.

ANOVA conducted on the raw attribute intensity ratings for the panel. The ANOVA was used to assess whether there were statistically significant differences between samples for each attribute and to evaluate the effects of other variables (facets), such as Assessor and Repetition, on the ratings.

Table 4.6. Summary of TIM Rasch analysis, and ANOVA results on raw attribute scores, showing attribute contributions to sample differences.

Attributes	+Ve Logit			-Ve Logit	
	Orange Fl.	**Sweetness**	**Cocoa Fl.**	**Saltiness**	**Milky Fl.**
Rasch Model Results					
Attributes Logit Measure[1]	0.75	0.74	0.21	**-1.2**	**-0.5**
Attributes OUTFIT Mnsq[2]	1.23	0.95	0.95	0.97	0.91
Panel ANOVA[3]					
F_{Sample}	**14.12*****	**5.66****	**2.13**	**4.34***	**0.38**
$F_{Assessors}$	**5.71*****	**8.45*****	**8.11*****	**17.43*****	**13.10*****
$F_{Repetition}$	1.85	3.24*	0.84	8.10***	12.76***
$F_{Assessors\ X\ Samples}$	0.95	0.85	0.96	1.00	1.15
$F_{Assessors\ X\ Repetition}$	1.27	1.32*	1.49**	1.56**	1.25
$F_{Sample\ X\ Repetition}$	0.83	0.20	0.19	0.39	0.59

Attributes are arranged from left to right by decreasing OUTFIT Mnsq value and are differentiated based on whether they were located on the positive (+Ve logit > mean) or negative (-Ve logit < mean) side of the logit scale.

[1]Value of the location of an attribute on the Rasch logit scale. Negative (-Ve) logit values signify low-intensity attributes (below the mean), and attributes are more challenging to rate when they are not overfit.

[2] Outlier-sensitive mean squares for attributes indicating whether the discrimination of an attribute differs from the average discrimination of other attributes in the test. The acceptable range for this data is 0.89 -1.12 and values <0.89 (overfit) signify a non-discriminating attribute.

[3]ANOVA on raw scores; F-values with p-value levels of significance: <0.001***, <0.01**, <0.05*; measures with no superscript symbols >0.05.

Notably, none of the attributes overfit the Rasch model (i.e., OUTFIT Mnsq < 0.89), suggesting that they were perceived differently across the samples. Further investigations using F-values from the ANOVA results revealed which samples were significantly different. For sample (F_{Sample}), the attribute with the highest F-value and the most significant p-value was *Orange flavour*, suggesting that it had a significant impact on sample differences. This finding aligned with the Rasch analysis results,

where *Orange flavour*, the most intense attribute (highest logit value = 0.75), underfit the Rasch model with an OUTFIT Mnsq value of 1.23 implying that on average, *Orange flavour* had the most variable intensity ratings across the samples. This was followed by *Sweetness* and *Cocoa flavour* which had identical OUTFIT Mnsq values. However, low-intensity attributes, *Milky flavour* and *Saltiness* (with negative logit values) also had high OUTFIT Mnsq values and were flagged as more challenging for the panel to rate. Although the ratings did not underfit the model, assessors assigned the most varied ratings, as indicated by the higher F-value for assessors ($F_{Assessors}$), with *Milky flavour* identified as slightly less challenging compared to *Saltiness*. The greater variance amongst parameters within a facet, the higher the OUTFIT Mnsq value (Linacre, 1995).

A visual representation of the hierarchy of attribute contributions to product differences is shown in **Figure 4.6**. This provides a clear and high-level basis for identifying key drivers of product differences, enabling analysts to target reformulation or quality control efforts accordingly.

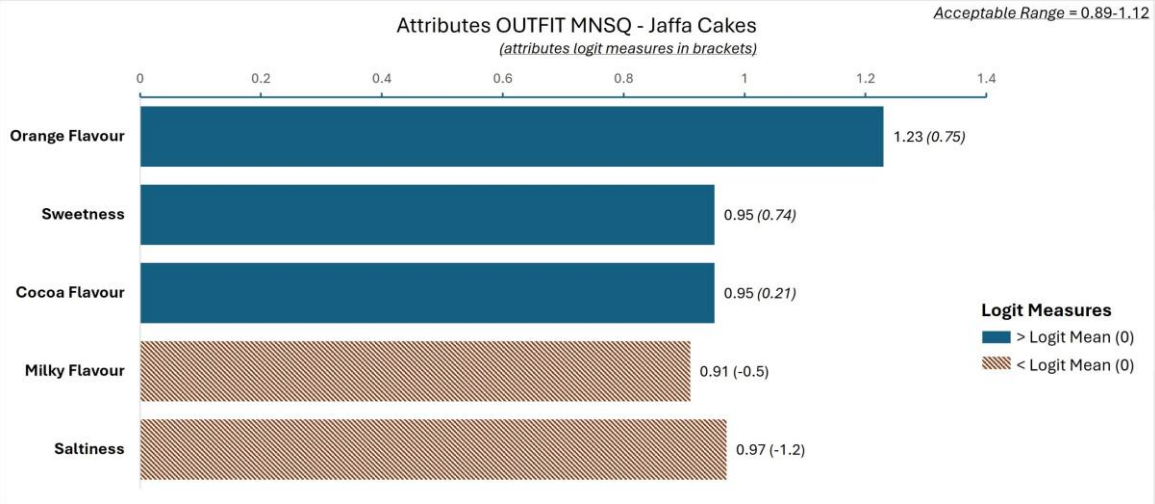


Figure 4.6. Sensory attribute contribution to overall differences between Jaffa cake samples based on Rasch logit measures (in brackets) and residual fit statistic (OUTFIT Mnsq). Attributes are colour-coded by logit sign: blue fill = positive logits (higher intensity, contributing more to product differences); red textured fill = negative logits (lower intensity, rated more inconsistently). This division helps distinguish attributes driving sample differences from those that were more challenging for the panel to rate.

When further insights such as assessing statistical significance or exploring specific attribute interactions, this information can be complemented with additional analyses like pairwise comparison tests and visualisation plots. It also supports decisions about whether the panel requires further training on more

challenging attributes or if certain attributes contribute little, if at all, to product differentiation and can be excluded from further analysis.

From the ANOVA results (**Table 4.6**) *Sweetness* had a significant impact on sample differences, while *Cocoa flavour* did not. However, Rasch results showed that these two attributes had the same OUTFIT Mnsq, indicating similar levels of inconsistency in how assessors rated them across the samples. This suggests that, although *Sweetness* was rated as more intense, both attributes exhibited comparable response irregularities possibly due to perceptual interaction or contextual influences. A likely explanation is the cognitive bias known as logical error (Myford & Wolfe, 2003; Kemp et al., 2018; Meilgaard et al., 2025), where assessors associate certain product characteristics in their minds, such as sweetness and chocolate (cocoa flavour), leading them to rate these attributes in a similar way despite their differences in average ratings.

Based on the Rasch analysis results, *Orange flavour* emerges as the primary contributor to the differences between the Jaffa cake samples. Upon reviewing the sample composition (Appendix **B.1**), it was found that the orange flavouring used in Brand A differed from that of Brand B and the Control, which both used the same flavouring. However, this conclusion was based solely on the information on the back-of-pack labels, and it is possible that the concentrations of the flavourings varied within Brand B and the Control. For salt content which had the most inconsistent ratings, the Control sample contained 1.9g, while Brands A and B contained 0.27g and 0.2g, respectively. While this might have constituted a perceptible difference at higher concentrations, the low intensity ratings for *Saltiness* suggest otherwise. Low-intensity attributes can be challenging to detect as they may be close to or below the threshold of detection for assessors (Lawless & Heymann, 2010; Meilgaard et al., 2015). Therefore, assessors may have struggled to rate them on the scale, and the variation in their ratings may have resulted from uncertainty about whether they were truly perceiving the attributes, hence the inconsistent ratings across repeated sessions and highly significant F-values for Repetition ($F_{Repetition}$). Discrepancies in the rating of *Saltiness* could have contributed to additional differences between the samples. While removing some inconsistent assessors could provide clarification, this was beyond the scope of the current study.

Since the TIM approach required assessors to focus on specific attributes, the results indicated that *Orange flavour* was the primary driver of the sample differences. As such, it should be the focus of any sample reformulation for product development or quality control, while assessors may need training on rating *Saltiness* and *Milky flavour* if these are to remain as key differentiators for the products, particularly when using a trained panel. Detecting and quantifying low-intensity attributes often requires extensive training (Lawless & Heymann, 2010; DLG, 2020). Although higher-intensity attributes are generally easier to perceive, this does not necessarily mean that assessors will rate them consistently. The results suggest that there were likely some erratic ratings from the untrained panel even for easily perceived attributes like *Orange flavour* and *Sweetness*, not only for low-intensity ones like *Saltiness*. This indicates that inconsistency may arise not only because some attributes are difficult to detect, but also because assessors interpret or apply attribute definitions differently. Therefore, targeted training to improve assessors' understanding of the attributes and where they lie on the scale remains relevant, even for attributes that are easily perceived.

4.4 Limitations of the study

Poor attribute representation

This study compared overall difference results from the DFC and those from Rasch-combined taste/flavour attribute ratings. However, the selection of the 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 flavour/taste attributes in rating the Jaffa cake samples. The former was the case for Brand A, where assessors' familiarity and possibly its appearance, led to it being rated as much more significantly different from the control, with the magnitude of the difference larger than that found in the TIM (corresponding Dunn's test p-values for TIM1 and DFCM1 = 4.54e-05 and 1.31e-33 respectively).

In contrast, the AR test focused solely on selected taste/flavour attributes, leaving potential variations in other sensory characteristics unaccounted for. As a result, the Total Intensity Measure (TIM) was estimated based only on these attributes. This narrow focus may have increased the risk of a Type I error in the TIM approach, where differences are identified that might not fully represent the overall product perception. While earlier analysis using the Friedman test yielded stronger significant differences, the current Kruskal Wallis results still identified differences, albeit to a lower degree. Similarly, the narrow focus could have led to a Type II error in the DFC, where assessors may have missed meaningful differences in the samples by focusing on the most prominent attribute difference from the control 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 results.

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. The study discussed in **Chapter 6** attempts to address this limitation.

Lack of instrumental analysis to verify product characteristics

A further limitation of this study is the absence of instrumental or analytical verification of the Jaffa cake samples' sensory attribute profiles. Sample selection relied on subjective informal tasting sessions and ingredient label information rather than instrumental confirmation that the samples differed in the selected attributes. The possibility that samples varied in other sensory dimensions, such as sponge texture, jam consistency, chocolate coating thickness, or secondary flavour notes, cannot be ruled out.

Although the target attributes for the AR test were primarily taste and flavour characteristics, assessors may have been influenced by texture-flavour interactions, where the physical structure of the product could be modulating the

release, diffusion, and perception of flavour volatiles. As Brouwer et al. (2024) found, variations in a product's physical matrix, such as changes in viscosity, can significantly alter flavour intensity, illustrating how structural factors can mediate sensory experience. In the present study, similar effects may have arisen from differences in sponge thickness or in jam gel consistency or chocolate coating properties, each of which could influence flavour release during consumption. These interactions align with the broader concept of the food matrix effect described by Aguilera (2019), where the structure and composition of foods shape their sensory perception.

Ideally, instrumental verification including texture profile analysis, compositional analysis, and headspace GC-MS for volatile flavour compounds like orange flavour would complement sensory assessment to confirm samples varied exclusively in the target attributes.

Differences in test structure and presentation design

Differences in presentation design and scaling between the DFC and AR tests represent another limitation. The DFC employed a comparative design in which assessors directly compared each sample with a physical reference. Together with the ordinal difference scale, this format not only heightens perceived differences but also provides an external scale anchor that simplifies judgments and supports consistent discrimination.

In contrast, the TIM used a monadic design combined with a category scale with anchors adapted from the Labelled Magnitude Scale (LMS), requiring assessors to rate each sample independently using internal references. Because the LMS is an absolute-intensity scale, assessors often interpret its verbal anchors literally, reserving extreme categories (e.g., "strongest imaginable") for stimuli perceived as unusually intense (Lawless & Heymann, 2010). This conservative response behaviour, together with the absence of a physical reference, may have compressed the effective scale range, and increased cognitive variability. Consequently, some of the observed differences in discrimination, strata, and reliability may reflect presentation and scaling-related effects rather than true analytical differences between the methods.

Ultimately, while the DFC served as the benchmark method because it is an established approach for measuring the magnitude of overall product differences,

it addresses a different sensory testing purpose than TIM. DFC confirms whether products differ overall, whereas TIM is diagnostic, identifying which specific attributes contribute to those differences. These distinctions in presentation design, scaling, and model structure should therefore be considered when interpreting the comparative results of this study.

4.5 Significance of the study

When Rasch measures of combined attributes (TIM) are subjected to univariate pairwise comparison tests, they reveal overall differences between products relative to a reference sample, like the DFC test. However, the sample requirements for the DFC are more demanding than for the AR tests used in the TIM approach, because each sample is evaluated in direct comparison to a reference. This effectively doubles the number of evaluations required for the DFC, especially when replicate assessments are included, as illustrated in **Figure B 2**. Consequently, it can be more resource-intensive, and assessors are more likely to experience sensory fatigue.

Additionally, while the DFC test is useful for quantifying the magnitude of perceived differences between samples, it only indicates that a difference of a certain magnitude exists without identifying which attributes drive that difference. Moreover, because assessors evaluate differences based on their own perceptions without specific guidance to what attributes to look for, there is a risk that irrelevant or unintended attributes may influence their assessments.

In contrast, the TIM method provides detailed, actionable insights that support decision-making in sensory quality control and product development. Targeted attributes can be included in the AR test, and Rasch measures of combined attributes can be used to compare individual test products or compare test products against a control using the appropriate post hoc tests. The DFC test, on the other hand, only allows for comparisons against a control and does not permit direct comparisons between individual test samples (Rogers, 2017). With the TIM method, the control sample can either be predetermined during the conceptualisation phase or selected retrospectively. Additionally, an action standard can be established to guide decisions on implementing product changes

and to identify areas where further investigation is needed to determine which attributes are significantly different. Researchers like Næs et al. (2010) have expressed that the margin between a significant and non-significant difference between samples may not always be clearly reflected by p-values, and significant p-values do not always translate into commercial importance (Kemp et al., 2018).

Additionally:

- The MFRM converts ordinal scores into interval-scaled measures and can separate the effects of multiple influencing variables (facets), resulting in a fairer and more accurate assessment of sample differences. Compared to conventional analysis methods (ANOVA / MANOVA) that assume data are interval-scaled, the MFRM provides a more robust evaluation of product differences. These insights are visually represented on an easily interpretable Wright map.
- The position of attributes on the logit scale reflects the hierarchy of dominant attributes perceived across all samples.
- Rasch model fit statistics (such as OUTFIT Mnsq) provide insights into which attributes most influence these differences. This helps analysts prioritise attributes for reformulation or quality control troubleshooting.
- The combination of OUTFIT Mnsq and the logit scale position reveals which attributes were easier or more challenging for the panel to evaluate, helping panel leaders identify which attributes may require additional assessor training.
- These results on attribute rating variations across samples can also be visually represented in easily interpretable plots, as shown in **Figure 4.6** above.

In the present study, attribute contributions from Rasch analysis were compared with ANOVA panel results, the standard statistical approach for evaluating assessor performance during selection, training, and descriptive analysis (Stone et al., 2012; Kemp et al., 2018; Meilgaard et al., 2025). However, no method currently integrates attribute intensity results into a single measure to quantify product differences while accounting for the effects of other variables (like assessor bias, and inconsistencies across replicates). Even Principal Component Analysis (PCA), at best, combines attributes into fewer dimensions, providing an overview of product differences, but it lacks the specificity that the MFRM-based TIM method

offers, including the ability to consolidate these results into a single measure of overall difference. Meanwhile, the MANOVA has limited diagnostics capacity, as it only indicates that samples differ based on combined attributes, without revealing which attributes are driving those differences. The TIM approach, in contrast, allows for the identification of which specific attributes are primarily driving product differences, consolidating these results into a single, comprehensive measure.

Furthermore, the MFRM approach streamlines sensory evaluation by also providing diagnostic details for assessors, using residual fit statistics for the *Assessor facet* and other quality control features. These insights help evaluate both panel and individual assessor performance and will be explored further in the next chapter. However, while MFRM offers clear analytical and interpretive advantages, its application requires specialised statistical expertise and access to dedicated software packages. These requirements may limit its broader adoption, particularly in routine industrial contexts where such resources or expertise may not be readily available.

The TIM approach relies on the use of pre-selected sensory attributes, meaning that the validity and completeness of conclusions depend on the adequacy of the chosen set of attributes. As shown in this chapter, omission of key attributes can lead to potentially misleading interpretations. To address this limitation, systematic attribute generation and selection procedures should be applied, such as preliminary profiling or descriptive analysis, to identify all relevant attributes likely to vary due to process changes, ingredient modifications, or product lifecycle stages. Careful attention to these factors ensures robust application across research and industrial contexts.

Nonetheless, the TIM approach provides some advantages over overall difference tests like the DFC. By directing assessors' attention to specific, relevant sensory attributes, TIM helps ensure ratings focus on the intended characteristics of interest. In contrast, holistic DFC ratings can be influenced by unrelated factors such as brand familiarity, or other incidental differences in presentation, even when efforts are made to control these factors. By constraining assessors to rate predefined attributes, TIM reduces the risk that such irrelevant or non-critical characteristics will confound judgments about product differences.

The following chapter builds on these insights by applying the MFRM to evaluate panel and assessor performance in both trained and untrained panels.

Chapter 5

Monitoring Panel and Assessor Performance with the Many-Facet Rasch Model (MFRM): A Comparison of Trained and Untrained Panels

5.1 Overview

In sensory quality programmes, the ability of trained assessors to detect product differences accurately is crucial for making informed research and development (R&D) decisions (Stone et al., 2012). This means that decisions are based on reliable evidence about product differences provided by these assessors, and inaccuracies can, for instance, lead to launching products with undetected flaws, or failing to identify a reduction in product quality due to ingredient substitutions. However, despite the level of training, individual differences continue to challenge the consistency and reliability of sensory difference and descriptive tests. Variability in perception, driven by factors such as culture, environment, experience, genetics, and personal preferences remains a persistent source of measurement variation (Næs et al., 2010; Meilgaard et al., 2025).

Additionally, systematic differences may arise during the sensory evaluation, further contributing to measurement variation. For instance, assessors may use rating scales differently, either limiting their responses to a narrow range of the scale, or consistently placing ratings at the higher or lower ends of the scale (Næs, 1990; Romano et al., 2008). This underscores the need for assessor performance checks, continuous panel monitoring and panel proficiency testing (Kemp et al., 2018) to ensure reliable and consistent results. Tomic et al. (2007) suggested combining several visualisation techniques, such as eggshell and correlation plots, to examine individual and panel performance, approaches that remain widely used today. However, they emphasised that methods to compensate for rater drift, as well as for level and range effects among assessors, are still lacking and would be highly beneficial for improving panel reliability.

While it is ideal for assessors to function as a machine, giving ratings in the same way, this is unrealistic. Sensory assessments rely on human judgment, so some

level of subjectivity is unavoidable, even with well-trained panels. Assessors may interpret sensory attributes in slightly different ways or become fatigued or distracted during evaluations (as discussed in section **2.3.2: Individual differences in sensory evaluation**). Traditional approaches that rely on consensus scoring assume that perfect agreements are both possible and necessary. When this expectation is not met, it raises concerns about the reliability and validity of the sensory data (Kemp et al., 2009; Raithatha & Rogers, 2018). The Rasch model does not require perfect agreement among assessors but rather expects consistency within individual assessors (internal consistency) in terms of the use and understanding of the rating scale (Linacre, 1994). Each assessor's ratings are treated individually, and their tendency to rate higher or lower compared to the rest of the panel is accounted for in the model (assessor severity). By simultaneously estimating both attribute intensities and assessor severity, Rasch analysis enables fairer comparisons on samples, across assessors with different standards without requiring extensive training on the uniform use of scales (as discussed in section **3.1.2: The Many-Facet Rasch Model (MFRM)**). Additionally, the model converts ordinal sensory ratings into interval-scale data, enabling the use of simpler 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.

Rasch model fit statistics for the *Assessor facet* detect rater effects and idiosyncrasies in individual ratings, identifying assessors whose scoring patterns deviate from the model's expectations relative to the rest of the panel. The model's separation statistics also provide insights into panel agreement, and overall panel reliability similar to conventional panel performance criteria as discussed later in this chapter (section **5.3.3: Comparison of trained and untrained panel performance**). In the conventional approach, several statistical techniques are available to monitor assessor performance, including univariate (e.g., ANOVA) and multivariate approaches (e.g., PCA). However, the ANOVA method requires separate analyses for each attribute and each assessor, which can be cumbersome and provides only a small fraction of the diagnostic information needed for a comprehensive evaluation. While multivariate methods, though useful for data

reduction, can be challenging to interpret. In contrast, Rasch analysis integrates both product differences and assessor performance within a single framework, offering a more direct overview of panel agreement and individual performance issues. These insights can be further complemented by traditional statistical tests to obtain a more nuanced understanding of assessor and panel sensory data.

Furthermore, the Rasch model subset linking capabilities (Linacre, 2012b; Engelhard & Wind, 2018; Andrich & Marais, 2019) enable comparisons across different groups or datasets over time. Differential Facet Functioning (DFF) (as discussed in section **2.5.2**) provides insights into systematic biases or group-related differences in product sensory assessments. This approach is particularly valuable for panel proficiency studies (monitoring assessor or panel performance overtime), as well as for understanding how sensory attributes contribute to product differences across cultures in global panels, or between trained assessors and target consumer panels with varying levels of expertise. Although DFF was not explored in this study, it presents a promising direction for future research.

This chapter explores the use of the MFRM for assessor performance evaluations and compares trained and untrained panel attribute intensity ratings on chocolate spread samples.

5.1.1 Objectives

The aim of this study was to explore the potential of the Many-Facet Rasch Model (MFRM) in examining panel and assessor performance.

The specific objectives were:

1. To compare the performance of trained and untrained panels in rating attributes intensities of chocolate orange spreads.
2. To examine the implications of Rasch assessor fit statistics for standard assessor performance criteria.
3. To identify untrained assessors whose performance is comparable to trained assessors, and to compare their results with the trained panel using Rasch analysis.

5.1.2 Study highlights

- Rasch group-level statistic (fixed chi-square for the *Assessor facet*) indicated generally consistent scale use for the trained panel, but greater inconsistencies for the untrained panel.
- The trained panel demonstrated a greater ability to discriminate reliably between the chocolate spread samples, whereas the untrained panel did not.
- OUTFIT mean square ranges revealed specific rating effects in both panels, aligning with raw rating scores observed in trellis plots*.
- The MFRM Wright map revealed differences in how each panel interpreted and rated the sensory attributes, indicating variability in scale use between the trained and untrained assessors.
- OUTFIT mean square and logit values revealed which sensory attributes were consistently assessed by each panel and which attributes posed greater challenges for them to rate reliably.
- PCAR revealed response dependency between attributes, driven by the presence of milk chocolate as an ingredient.

5.2 Sensory study: materials and methods

Data for this study were obtained from an existing dataset (Gill et al., 2024).

5.2.1 Samples

Chocolate spread was chosen for this study, as the trained panel from the global chocolate manufacturing company had prior experience evaluating chocolate products, though not specifically chocolate spreads. Three brands of chocolate spread were selected, based on noticeable differences in orange flavour and sugar content, as indicated on the back-of-pack labels. The three brands, purchased from UK retail stores, consisted of two chocolate orange spreads and one chocolate spread without orange flavouring. In terms of *Sweetness*, one sample contained

* Trellis plots are multi-panel charts used here to show each assessor's raw score distributions for each attribute, product, and replicate, helping to visualise variation and potential rating effects across the panel.

maltitol, a sugar replacer, while the others used sucrose. The sample composition is provided in **Table B 2**.

Plain white bread was used as a "carrier" (Lawless & Heymann, 2010) for all three samples, with 5g of each chocolate spread applied to one side of a rectangular slice measuring 2.5cm x 4cm. This approach was necessary to represent the typical context in which chocolate spreads are consumed, thereby minimising potential psychological biases. Samples were stored at room temperature (20±3˚C) in their original packaging until they were ready to be presented.

5.2.2 Participants

Ethical approval for the sensory study was granted by the MaPS and Engineering Joint Faculty Research Ethics Committee at the University of Leeds (Appendix **A.2**).

Participants were comprised of a trained (n=7) and an untrained panel (n=24), as mentioned previously. Minimal demographic information was provided; however, it was noted that the trained panel, loaned by the global chocolate manufacturing company, were all females with at least 2 years of experience evaluating the sensory profile of chocolate products. The untrained panel comprised students from the University of Leeds, who participated voluntarily and received no incentive.

Each participant received detailed information about the study requirements, as well as the data protection and sharing disclaimer, and was required to sign consent forms before commencing the study.

5.2.3 Panel training

As the trained panel was already quite experienced in evaluating chocolate bars, there were only two training sessions conducted on separate days. The first session familiarised the assessors with the chocolate spread samples, methodology, and the modified LMS rating scale (described in **5.2.4**), which differed from the unstructured line scales typically used for Qualitative Descriptive Analysis (QDA) (discussed in section **2.4.2**: Rating scales), which they were already familiar with. The second session focused on generating descriptor terms for the product. During this session, they identified five attributes: three flavour attributes (*Orange*, *Milky*, and *Cocoa*) and two taste attributes (*Sweetness* and *Saltiness*) from their

evaluation of a chocolate orange spread sample with similar nutritional and flavour characteristics to those intended for the study. This approach was necessary to prevent sensory bias due to overexposure and memory effects (Meilgaard et al., 2015), as using the same samples for both the preliminary session and the main evaluations could influence the panel's perception. To ensure a focused lexicon, assessors were instructed to concentrate on the taste and flavour characteristics of the products while disregarding any textural differences.

5.2.4 Sensory evaluation procedures

In both panels, a Randomised Complete Block Design (RCBD) with counterbalancing was used for the experiments to minimise potential biases, such as carryover and order effects, and samples were evaluated in triplicates. The trained panel completed their replicate sessions over two days due to time constraints, with the first two sessions held on the first day and the third session on the following day. The untrained panel, on the other hand, completed all replicate sessions on the same day. Samples were presented monadically, at room temperature (20±3˚C), on 15 cm white paper plates, each labelled with random 3-digit codes.

The assessors rated the perceived intensities of five taste/flavour attributes: *Orange flavour*, *Sweetness*, *Cocoa flavour*, *Milky flavour*, and *Saltiness*, which were generated by the trained panel. They were instructed to taste each sample, focusing on the specified attributes, chew and swallow the bread (carrier) with the chocolate spread sample, and then proceed to the next page of the questionnaire. On this page, they were asked to rate the strength of each attribute in the sample they had just tasted. All five attributes were presented on the same page of the questionnaire with an additional comment section for any other impressions about the sample.

The same 8-point categorical intensity scale, as used in Chapter 4 was employed in this study. It ranged from 0 to 7 with labels adapted from the Labelled Magnitude Scale (LMS) (Green et al., 1996). 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. A copy of the questionnaire can be found in Appendix **C.2**.

Assessors were provided with some water to cleanse their palate between sample evaluations.

5.2.5 Data analysis

As with the previous chapter, Rasch and statistical analyses were according to the procedures described in section **3.3**. The attribute intensity ratings (AR) data from each panel were fitted separately to the MFRM; that is, one model for the trained panel and another for the untrained panel, each including four facets (*Assessors, Samples, Repetition* and *Attributes*), covering the variables under study. The Rasch model used in this study is the same as ***...Equation 4.1***. and is outlined below.

$$\boldsymbol{ln}\left(\boldsymbol{P_{mnrik}}/\boldsymbol{P_{mnrik-1}}\right) = \boldsymbol{\beta_m} - \boldsymbol{\theta_n} - \boldsymbol{\rho_r} - \boldsymbol{\delta_i} - \boldsymbol{\tau_k}$$

*...Equation **4.1***

Where:

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

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

β_m= degree of leniency or severity of assessor (m) in rating attribute intensities

θ_n= degree of difference in the total intensity measure for sample (n)

ρ_r = degree of difference between ratings of samples in a replicate session (r)

δ_i = the average degree of intensity of sensory attribute (i) across the samples

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

5.2.5.1 Panel Performance Evaluation

Rasch analysis and ANOVA techniques were used to evaluate individual and panel performance, based on standard performance indices (discrimination, panel agreement and repeatability) as described later in the chapter. Insights from both methods were compared to highlight the strengths and limitations of each approach, and results from the trained and untrained panels were also examined.

5.2.5.2 Convergence analysis

Convergence analysis was conducted to examine how panel size influenced product discrimination ability in the Many-Facet Rasch Model (MFRM). To assess the stability of the convergence patterns, the sampling procedure was repeated across two iterations, each using a different random draw of assessor subsets. Random subsets of assessors at panel sizes n = 7, 10, 12, 15, 18, 21, and 24 were generated from both the untrained panel (n = 24) and trained panel (n = 7) using random sampling without replacement in RStudio, with a fixed random seed (set.seed(123)) within each iteration to ensure reproducibility.

To enable comparison across equivalent panel sizes, the trained panel data were expanded by duplicating the ratings from the 7 assessors to create a pool of 24 simulated assessors, from which subsets were then randomly sampled using the same procedure as for the untrained panel. Rasch analyses were conducted in FACETS for each subset, and the resulting fixed chi-square statistics for the product (*Sample facet*) were extracted as indicators of product differentiation. Convergence plots were generated to visualise the relationship between panel size and discrimination ability for both panel types.

5.3 Results and Discussion

5.3.1 Fit of data to the Many-Facet Rasch Model (MFRM)

As in the previous chapter (section **4.3.1**), Rasch model fit statistics were examined for both panels to assess whether their data met the model assumptions for unidimensional measurements. The results of the global model fit, *Assessor facet* fit statistics and response dependency checks are presented in **Table 5.1**.

An acceptable global model fit of the data is when about 5% or less of absolute standardised residuals is ≥2, and about 1% or less is ≥3 (Linacre, 2022; Eckes, 2023). All models showed an acceptable global model fit suggesting that overall, the data in each model aligns with the assumptions of the Rasch model and there are no major inconsistencies that may distort the measurement. Only the individual fit statistics for the *Assessor facet* are presented, as assessors were the focus of this study. However, all other facets showed a 100% fit to the models, except for the

Attributes facet, where *Orange flavour* showed underfit in both panels. This will be discussed further in section **5.3.3.3.1**.

Table 5.1. Summary of Rasch model fit and assessor fit indices for the trained and untrained panels.

Criteria	Trained Panel (n=7)	Untrained Panel (n=24)
Global Fit (%StRes)[1]		
≤ 5% ≥ 2	5.0 (16)	2.8 (30)
≤ 1% ≥ 3	0.3 (1)	0.1 (1)
Total count[2]	**315**	**1080**
Assessor Fit (OUTFIT Mnsq, Nr =45)[3]		
% Fit (0.57-1.42)	**71.4 (5)**	**63.0 (15)**
%Overfit (≤ 0.57)	14.3 (1)	21.0 (5)
% Underfit (≥1.42)	14.3 (1)	12.0 (3)
% Extreme Misfit (>2.0)	0.0 (0)	4.0 (1)
Unidimensionality[4]		
1st contrast eigenvalue (<2)	1.81	1.96
LID (Corr. of StRes <0.3)[5]		
Sweetness-Milky flavour	0.13	0.31
Cocoa flavour - Milky flavour	0.09	NA[6]

[1] Percentage (number of observations in brackets) of absolute standardised residuals (StRes).

[2] Total number of observations used for the estimation of the Rasch model parameters.

[3] Outlier-sensitive measure of unweighted mean squares indicating deviation of the *Assessor facet* estimates from Rasch model predictions. The acceptable fit range (0.57-1.42) was determined using $1 \pm 2\sqrt{(2/Nr)}$ (Wu & Adams, 2013; Eckes, 2023), where Nr is the number of responses used for parameter estimation.

[4] Eigenvalue of the unexplained variance in the first contrast, not accounted for by the Rasch model, in PCAR.

[5] Local Item Dependency (LID) examined through the correlation of standardised residuals (Corr. of StRes) between attributes, with values > 0.3 indicating that items (attributes) are dependent.

[6] NA =Not applicable meaning that the attributes were not flagged as potentially dependent for the panel.

The acceptable range for the OUTFIT Mnsq for assessors was calculated following Eckes (2023), based on the number of responses per assessor in each panel (Nr=45), as discussed in section **3.3.1.6**. Each assessor had an equal number of responses, having rated the five attributes across three samples in three replicates (5 x 3 x 3 = 45). Although Linacre (2024b) and (2025b) suggests a rule of thumb for an acceptable mean square fit statistics, with lower and upper limits of 0.5 and 1.5

respectively, several researchers (Myford & Wolfe, 2009; Wu & Adams, 2013; Engelhard & Wind, 2018; Bond et al., 2020; Eckes, 2023; Linacre, 2025b) have shown that these critical ranges can vary and should be determined based on the specific assessment context and sample size.

Unidimensionality and Local Item Dependence (LID) were examined by Principal Component Analysis of Residuals (PCAR). Unidimensionality is confirmed when the eigenvalue of the unexplained variance in the first contrast is <2, and Local item dependence (LID) is identified when the residual correlation between two attributes exceeds 0.3 (Ramp et al., 2009; Christensen et al., 2017).

In both panels, the unexplained variance in the first contrast (representing residuals in the largest secondary dimension) had an eigenvalue of 1.81 and 1.96 for the trained and untrained panels, respectively. This indicated a strength of 2 out of 5 attributes, suggesting the possibility of a secondary dimension. The standardised residuals correlation for the attributes in the trained panel indicated no dependent attributes, while in the untrained panel, a correlation of 0.31 was observed between *Sweetness* and *Milky flavour*, suggesting a potential local dependency. Although this value barely exceeded the typical threshold of 0.3, it warranted further investigation into possible underlying causes, as discussed later in the chapter.

5.3.2 Representing the Overall Difference Construct

Wright maps for the trained and untrained panels are presented in **Figure 5.1** and **Figure 5.2** respectively, with all four facets (*Assessors*, *Samples*, *Repetition*, and *Attributes*) positively oriented, as described in previous chapters. The *Sample facet* was non-centred, while the other facets were centred at the mean (0 on the logit scale) to serve as a reference point. Consequently, sample locations were adjusted by considering the severity of assessors, the average intensity of attributes, and the intensity ratings in repeated sessions, representing the *Assessor, Attribute,* and *Repetition facets*, respectively. In the *Assessor facet*, assessors with higher logit values are more lenient, generally assigning higher scores on the rating scale; in the *Sample facet*, samples with higher logit values have higher Total Intensity Measure (TIM); in the *Repetition facet*, replicate sessions where higher intensity ratings were

assigned on average have higher logit values; and in the *Attribute facet*, attributes with higher average intensity ratings have higher logit values.

5.3.2.1 Trained Panel Representation

The Wright map in **Figure 5.1** below provides an overview of the trained panel's ratings of the overall difference between the samples.

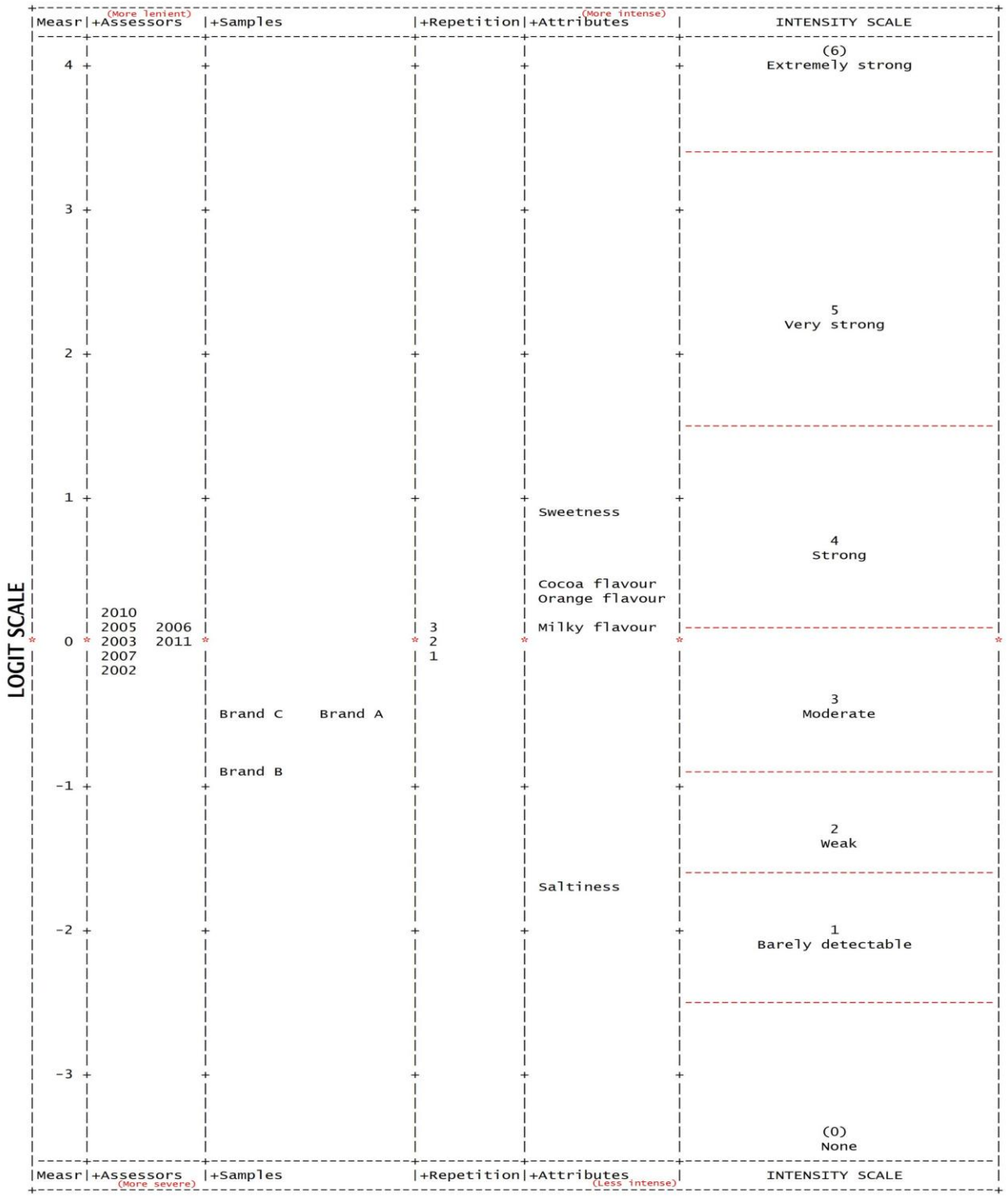


Figure 5.1. Trained Panel Many-Facet Wright Map.
From left to right, the columns represent: Rasch model measures on the logit scale (Measr); the *Assessor facet*, showing 7 assessors (IDs 2002-2011); the *Sample facet*, displaying Brands A-C; the Repetition facet, indicating replicates 1-3; the *Attribute facet*, listing the 5 attributes; and finally, the AR intensity rating scale, with horizontal lines marking half-point thresholds where the probability of assigning a higher rating exceeds that of assigning a lower adjacent rating.

In the Assessors facet, all assessors were distributed around the mean within a narrow range (−0.2 to 0.2 logits), suggesting that after accounting for measurement error (such as inconsistent replicate ratings), they used similar parts of the scale to assign ratings.

For the *Sample* and *Repetition facets*, the combined attribute intensity ratings across each sample were below the average (0 on the logit scale). The *Repetition facet* showed that differences across the three replicated sessions were not significant, indicating that the assessors most likely rated the samples consistently across sessions. Samples positioned higher on the scale were perceived to have greater intensity, based on ratings averaged across all attributes.

The latent variable "*Overall Difference*" is reflected in the Total Intensity Measure (TIM), which is represented by the location of the samples on the logit scale. Brands A and C were located very close together, with logit measures of -0.51 and -0.54 respectively, whereas Brand B was positioned much lower at -0.86 logits, indicating a noticeable difference from the others. The standard error for all three samples was 0.10. Given that Brand B's difference from Brands A and C is approximately three times the standard error, this suggests a potentially significant difference. Rasch separation statistics will reveal whether the observed difference is statistically significant, while TIM values will be used in multiple comparison tests to determine how much specific samples differed from one another.

The *Attribute facet* and intensity scale showed that *Sweetness* was the most dominant attribute (i.e., the most intense or easiest perceived attribute across the samples). *Orange flavour* and *Cocoa flavour* were perceived as strong, while *Milky flavour* was positioned at the threshold between moderate and strong, indicating it was generally perceived as strong, since the probability of a strong rating has exceeded that of moderate. *Saltiness*, however, was positioned at the barely detectable level, suggesting that assessors generally gave it the lowest ratings, which averaged within this range across samples, making it the least intense attribute (see **Table 5.4)**. The OUTFIT Mnsq for individual attributes will reveal which attributes are driving the differences between the samples, as discussed later in the chapter. Additionally, the intensity scale revealed redundant scale categories, as category 6 - Extremely Strong was barely used, while 7 - Strongest Imaginable Oral

Sensation was never used by the panel and did not appear on the Wright map (see **Appendix D: Rating Scale Category Statistics**).

5.3.2.2 Untrained Panel Representation

For the untrained panel represented in **Figure 5.2** below, two assessors appeared to be using different parts of the scale.

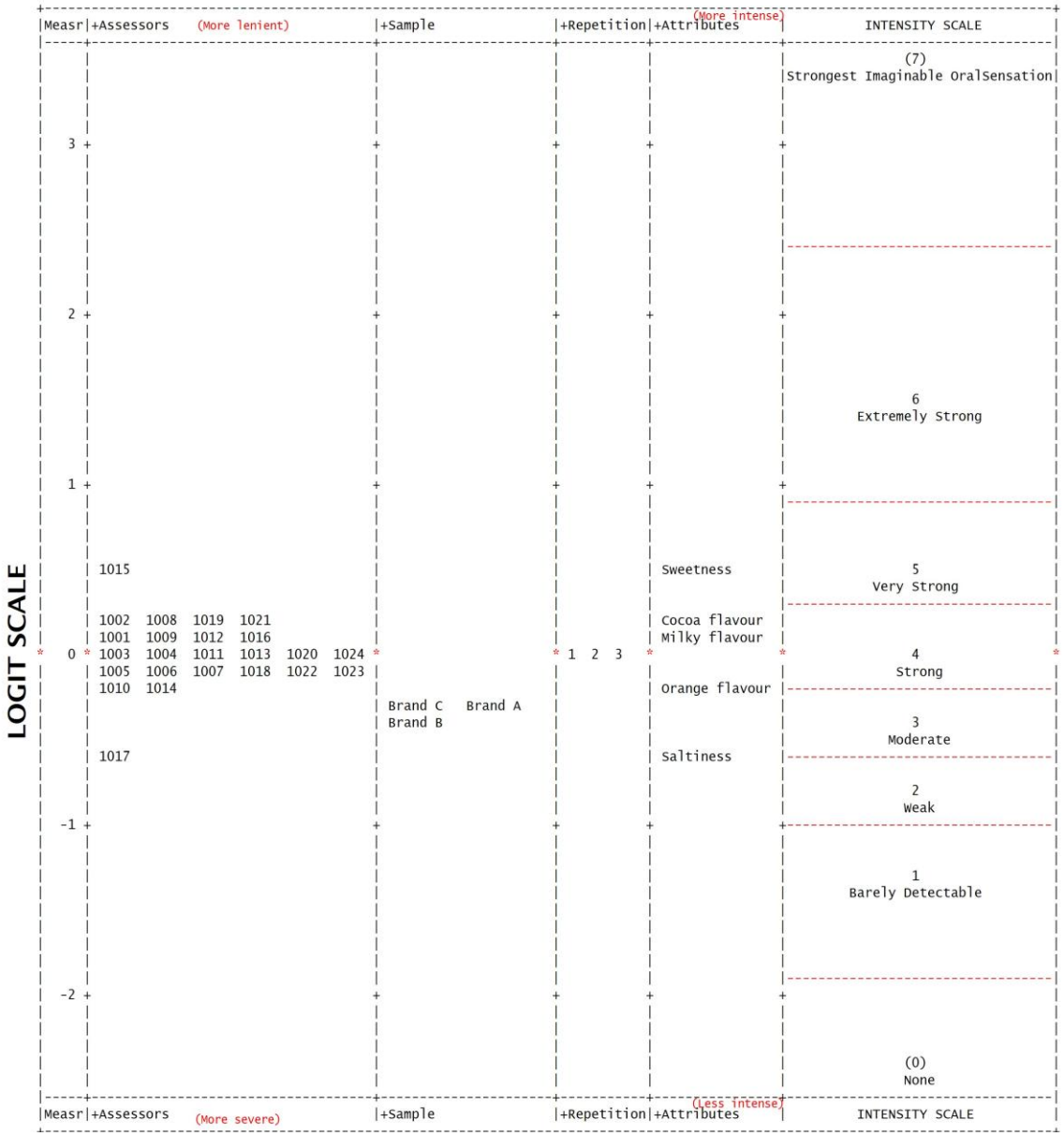


Figure 5.2. Untrained Panel Many-Facet Wright Map.
From left to right, the columns represent: Rasch model measures on the logit scale (Measr); the *Assessor facet*, showing 24 assessors (IDs 1001-1024); the Sample facet, displaying Brands A-C; the Repetition facet, indicating replicates 1-3; the *Attribute facet*, listing the 5 attributes; and finally, the AR intensity rating scale, with horizontal lines marking half-point thresholds where the probability of assigning a higher rating exceeds that of assigning a lower adjacent rating.

Assessor 1015 was the most lenient, consistently assigning higher ratings to the samples, while Assessor 1017 was the most severe, consistently assigning lower ratings on average. Their different rating behaviours will be flagged in the Assessor OUTFIT analysis, discussed later in the chapter. Overall, assessors were distributed around the mean within a narrow range (−0.2 to 0.2 logits), similar to the trained panel, suggesting that most assessors used similar parts of the scale to assign ratings. Attribute intensity ratings for the samples were also generally below average on the logit scale, and the averaged ratings were consistent across replicate evaluations. In contrast to the trained panel, although the relative positioning of the samples was consistent, with Brand B rated lower than the others, the average ratings did not differ significantly as later confirmed by the separation statistics (**Table 5.3**).

The *Attribute facet* and intensity scale revealed that *Sweetness* and *Saltiness* were again, the most and least intense attributes, respectively, consistent with the trained panel. However, the locations of *Milky flavour* and *Orange flavour* were reversed, with *Orange flavour* now positioned below average, raising the question of whether the trained panel had perceived *Milky flavour* as stronger than *Orange flavour* across the samples. This will be further investigated in the following discussions. The half-point thresholds were narrower, indicating that ratings were more evenly distributed across the scale categories, which reflects some imprecision in how products were rated. However, from category 6 (Extremely Strong) onward, the thresholds widened, likely due to the less frequent use of the highest categories (see **Appendix D: Rating Scale Category Statistics)**.

Rating scale category diagnostics are not discussed in this chapter, as the focus is primarily on examining individual and panel performance using insights from the OUTFIT mean square. However, as demonstrated in **Chapter 4**, the Many-Facet Rasch Model (MFRM) can also be used to assess the functionality of the rating scale and guide decisions regarding the need for scale revisions, which can be particularly beneficial when developing sensory quality programmes. “Hybrid models” (Myford & Wolfe, 2003) as described in **Table 2.2. Summary of Rasch Models**, can provide further insights into how each individual utilises the rating

scale. Summary statistics for rating scale functionality for both panels are provided in **Appendix D.**

5.3.3 Comparison of trained and untrained panel performance

The performance of the two panels were compared with respect to standard performance indices in ISO 11132:2021 (British Standards Institution, 2021), where performance is defined as *"the measure of the ability of a panel or an assessor to make reliable and valid attribute assessments across the products being evaluated"*. Kemp et al. (2009) define validity as the proximity of an assessors ratings to the average ratings of the panel, while Raithatha and Rogers (2018) broaden this to overall validity, referring to the extent to which sensory panel results can be reliably used to inform business action standards. The panel agreement, discriminatory ability, and repeatability (described in detail in the following sections) were assessed using both the Rasch model, which provides a holistic overview of the panel, and the conventional three-way ANOVA approach, which was conducted on the raw sensory scores, and analysed separately for each attribute under study. In both methods, the variables are treated as fixed effects; that is, as population model statistics within the Rasch framework, following Linacre (2025d), since the focus is specifically on these variables, and no generalisation beyond the observed data is intended. The results are summarised in **Table 5.2** and **Table 5.3**.

5.3.3.1 Panel agreement

This refers to the degree of alignment between assessors' average product scores as defined by ISO 1132:2021. It describes the ability of assessors within a panel to be consistent, demonstrating the same sample order where differences based on an attribute exist. The level of agreement should be sufficient for the panel mean to serve as a representative measure of the product differences (Raithatha & Rogers, 2018).

In the ANOVA, the panel lacks agreement when the **interaction factor between sample and assessor** ($F_{Assessors \times Samples}$ in **Table 5.2**) is significant ($p<0.05$). This suggests that some assessors differed in the relative ordering of samples, as implied by their assigned ratings for an attribute. This is illustrated in the trellis plots discussed

later in section **5.3.4**. According to IS0 1132:2021, the higher the number of "*key*" attributes with a significant interaction factor, the less consistent the panel is.

Key attributes are those that either show significant product discrimination by the entire panel or are associated with predefined differences between the samples. When inconsistencies are observed, further investigation at the individual assessor level is required, followed by appropriate corrective actions (e.g. assessor re-training).

Table 5.2 revealed that the trained panel consistently rated three out of five attributes (see $F_{Assessors \times Sample}$). The key attributes were *Milky flavour* and *Cocoa flavour*, while *Saltiness* was a non-key attribute as there was no significant difference between the samples (F_{Sample}). Notably, *Orange flavour* had the highest F-value ($F = 78.82$, $p<0.001$), indicating it as a key attribute contributing to product differences. However, a significant Assessor x Sample interaction ($F_{Assessors \times Sample}$) was also observed for this attribute. The interaction plots (**Figure 5.4)** revealed that while most assessors rated Brand B lowest, a few reversed the order for Brand A and Brand C, suggesting inconsistency in how the attribute was evaluated across assessors. This highlights the need for additional training to improve panel consistency and reliability. Similarly, *Sweetness*, which was slightly less significant ($p<0.05$), also showed some inconsistencies among assessors. These findings warrant further investigation, and individual ANOVAs for each assessor will be examined in the following section, to identify those who may need further training.

On the other hand, the untrained panel was not in agreement, as they were highly inconsistent in rating all the attributes, with the $F_{Assessors \times Sample}$ interaction effect showing significant differences ($p<0.01$) across all attributes.

From a Rasch model perspective, panel performance is evaluated holistically using the model's **separation statistics** (**Table 5.3** below), which are based on the average ratings assigned by the panel across all attributes, samples, and replicate evaluations.

The **fixed Chi-square (χ^2)** statistic is used as an indicator of panel agreement testing the hypothesis that, after accounting for measurement errors, the severity of all assessors is the same (Myford & Wolfe, 2004; Linacre, 2025c).

Table 5.2. Summary of trained and untrained panel Rasch analysis and ANOVA results on attribute contributions to sample differences.

Attributes	**Trained Panel (N=7)**					**Untrained Panel (N=24)**				
	+Ve Logit				**-Ve Logit**	**+Ve Logit**			**-Ve Logit**	
	Orange Fl.	Milky Fl.	Sweetness	Cocoa Fl.	Saltiness	Milky Fl.	Cocoa Fl.	Sweetness	Orange Fl.	Saltiness
Rasch Model Results										
Attributes Logit Measure[1]	0.26(0.13)	0.10(0.12)	0.91(0.14)	0.39(0.13)	**-1.66(0.14)**	0.06(0.04)	0.19(0.04)	0.47(0.05)	**-0.16(0.04)**	**-0.55(0.05)**
Attributes OUTFIT Mnsq[2]	1.92	0.73	0.65	0.47	1.22	1.05	0.83	0.71	1.60	0.77
Panel ANOVA[3]										
F_{Sample}	**78.82*****	**10.87*****	**3.65***	**6.44****	**1.27**	**188.34*****	**23.16*****	**92.27*****	**87.82*****	**3.78***
$F_{Assessors\ X\ Sample}$	**5.44*****	**1.14**	**3.24****	**2.07**	**1.59**	**2.58*****	**3.87*****	**4.98*****	**2.05****	**2.78*****
$F_{Assessors}$	10.02***	6.84***	21.53***	5.10**	25.92***	10.70***	8.27***	7.63***	2.93***	21.32***
$F_{Repetition}$	4.08*	1.67	3.65*	1.49	0.27	3.20*	0.52	2.77	2.27	0.91
$F_{Assessors\ X\ Repetition}$	2.64*	1.09	4.47***	0.97	1.05	1.67*	1.52*	2.66***	0.82	1.25
$F_{Sample\ X\ Repetition}$	3.29*	0.79	2.24	0.74	0.57	1.41	2.04	1.46	0.45	2.40

For both panels, attributes are arranged from left to right by decreasing OUTFIT Mnsq value and are differentiated based on whether they were located on the positive (+Ve logit > mean) or negative (-Ve logit < mean) side of the logit scale. N signifies the total number of assessors in a panel.

[1] Value of the location of an attribute on the Rasch logit scale: Negative (-Ve) logit values signify attributes with intensities below the mean (low intensity), while positive (+Ve) logit values signify attributes with intensities above the mean (high intensity). Standard errors (S.E) for each estimate are shown in brackets.

[2] Outlier-sensitive mean squares for attributes indicating whether an attribute's discrimination differs from the average. For the trained panel, the acceptable range is 0.64-1.36, with values <0.64 (overfit) signalling a non-discriminating attribute. For the untrained panel, the acceptable range is 0.81-1.19, with values <0.81 indicating non=discrimination.

[3] ANOVA on raw scores; F-values with p-value levels of significance: <0.001***, <0.01**, <0.05*; measures with no superscript symbols >0.05.

In this context, severity refers to an assessor's tendency to consistently assign higher (lenient) or lower (severe) intensity ratings across samples, relative to other assessors in the panel. The panel is not in agreement when χ^2 is significant ($p < 0.05$), indicating that at least two assessors have significantly different severity levels. The **Separation Index (Strata)** indicates how many statistically distinct levels of severity exist among the assessors. The separation **Reliability** reflects how precisely these differences in severity are measured, relative to the error in the estimates. The separation statistics for the *Sample facet* are used to assess the panel's ability to discriminate between the samples.

For the *Assessor facet*, reliability values closer to zero (0) are desirable, as they indicate that there is no statistical distinction between lenient and severe assessors, suggesting that, on average they rated the samples using similar parts of the scale. For the *Sample facet*, however, higher reliability values closer to one (1) are ideal, as they suggest greater discrimination between the samples by the panel.

From **Table 5.3**, Rasch model chi square (χ^2) for assessors in the trained panel was not significant, indicating that they exhibited the same severity level on average, reflecting the effectiveness of their training. Meanwhile, assessors in the untrained panel showed different severity levels after accounting for measurement errors, as indicated by the highly significant assessor chi-square ($p<0.001$). A Strata value of 2.87 and a reliability index of 0.78 indicate the presence of approximately three statistically distinct levels of assessor severity, as was revealed in the *Assessor facet* of the untrained panel Wright map (**Figure 5.2**).

The MFRM examines panel agreement in terms of the order in which products differences are ranked (as in the ANOVA), at an individual level, using the **Point-Biserial Measure correlation (PT measure)** also termed the **Single Rater – Rest of Raters (SR/ROR) correlation** (Myford & Wolfe, 2004). This reflects how assessors rank samples relative to other assessors in the panel and is further discussed in individual performance evaluations (section **5.3.4)**.

Table 5.3 . Comparison of trained and untrained panel based on Rasch Model Statistics.

Rasch Model Statistics[1]	**Trained Panel (n=7)**	**Untrained Panel (n=24)**
Panel Agreement (*Assessor facet*)		
Fixed $\chi^2_{Assessor}$ *(p>0.05)*	0.54	0.00×
Strata $_{Assessor}$	0.33	2.87
Reliability $_{Assessor}$	0.00	0.78
Panel Discrimination (*Sample facet*)		
Fixed χ^2_{Sample} *(p<0.05)*	0.02	0.46×
Strata $_{Sample}$	1.95	0.33
Reliability $_{Sample}$	0.59	0.00

[1] Rasch model separation statistics (with required criteria in brackets) corresponding with standard panel performance criteria (and the related facet in brackets). The null hypothesis (H_0) for the fixed chi-square test is that all elements within the facet are the same. Therefore, $p < 0.05$ indicates a statistically significant difference in the facet parameters.

× Panel performance criteria is unmet.

Likewise, the $F_{Assessor}$ main effect in the ANOVA reflects variations in ratings assigned by assessors on average, independent of samples (Stone et al., 2012), which is similar to the MFRM in its assessment of panel agreement in terms of rating severity levels. However, while the ANOVA (**Table 5.2**) found that assessor tendencies in using the rating scale differed across all attributes, the Rasch model did not. This is because the Rasch model offers a more precise measurement by considering individual rating patterns across all samples, attributes, and replicate evaluations, adjusting for how consistently assessors tend to rate with varying severity or leniency. After which it then models leftover inconsistencies, both between-group (main effects) and within-group (interaction effects) as unexplained variations, captures them as measurement errors, and flags them in the OUTFIT statistics (Linacre, 1995). What remains, then, is the true variance from the main effects, reflecting the real differences between the parameters in each facet (assessors, samples, and repetition). The OUTFIT Mnsq results for attributes and assessors, which illustrate how misfitting ratings were identified and addressed, are discussed in detail in sections **5.3.3.3.1** and **5.3.4**, respectively. Unlike ANOVA, which reports averaged differences across all assessors, the Rasch model provides these insights on a more granular, individual level.

5.3.3.2 Panel repeatability

This refers to the average degree of homogeneity between replicate assessments of the same product per assessor (ISO 1132:2021). In other words, it measures how consistently assessors evaluate the same products under similar test conditions, typically across replicate sessions. The Repetition main effect and interaction terms in the panel ANOVA ($F_{Repetition}$, $F_{Assessor\ x\ Repetition}$, and $F_{Sample\ x\ Repetition}$) offer a high-level indication of which attributes assessors rated inconsistently across replicate evaluations on average (Ho, 2015). The Repetition main effect ($F_{Repetition}$) tests for overall mean differences across replicate evaluations, the Assessor x Repetition interaction ($F_{Assessor\ x\ Repetition}$) checks if assessors differ in consistency in rating the samples, and the Sample x Repetition interaction ($F_{Sample\ x\ Repetition}$) checks if some samples are more consistently rated than others.

As shown in **Table 5.2**, replicate evaluations for some assessors ($F_{Assessor\ x\ Repetition}$) in the trained panel varied for *Orange flavour* and *Sweetness*. However, these variations only influenced the *Orange flavour* ratings of the samples ($F_{Sample\ x\ Repetition}$).

In the untrained panel, replicate ratings for *Milky flavour*, *Cocoa flavour*, and *Sweetness* varied within assessors, but these variations were not substantial enough to influence their overall sample ratings when averaged. This was also reflected in the respective panel Wright maps where replicate evaluations for the trained panel (**Figure 5.1**), ranged from 0.06 to -0.10 logits (SE = 0.10), showing greater dispersion across repeated ratings. In contrast, the untrained panel's replicate evaluations (**Figure 5.2**) clustered tightly between 0.01 and -0.01 logits (SE = 0.03), indicating higher repeatability in overall sample ratings.

For a more detailed assessment, repeatability can be estimated by analysing individual assessor response patterns, through the use of distribution plots (Stone et al., 2012).

This approach will be examined later in the chapter.

5.3.3.3 Panel discrimination

The panel discrimination measures the ability of a panel to significantly distinguish between products. It is indicated by a significant difference in sample means for an attribute ($p<0.05$) in the three-factor panel ANOVA (F_{Sample} in **Table 5.2**). When the samples are significantly different, post hoc multiple comparisons are conducted

to determine which specific samples differ from one another. However, as previously mentioned, these results are only reliable when ratings are generally consistent, and the analyst judges that there is sufficient agreement between assessors within the panel (Raithatha & Rogers, 2018). The ANOVA results revealed that for the untrained panel, all the attributes significantly differentiated the samples ($p<0.05$). However, this panel also showed inconsistency in ratings for all the attributes ($F_{Assessor \times Sample}$) and showed variation in their rating tendencies across all attributes ($F_{Assessor}$). Consequently, their results are not reliable. The Rasch separation statistics in **Table 5.3** above showed that differences between samples were not significant ($\alpha=0.05$), as the differences across attributes were not systematic enough to be considered meaningful.

Rasch analysis of the trained panel data revealed significant differences between the samples (**Table 5.3**), with a χ^2 p value of 0.02. A strata value of 1.95 indicated the presence of approximately two distinct sample groups. However, the separation reliability was relatively low at 0.59, suggesting that a portion of the observed differences may have been due to measurement error. In the trained panel ANOVA (**Table 5.2**), rating tendencies differed across all attributes ($F_{Assessor}$) similar to the untrained panel, indicating individual differences in scale use despite assessor training. Additionally, four out of five attributes (except for *Saltiness*) were significantly different ($p<0.05$). However, two of these attributes, *Orange flavour* and *Sweetness* were rated inconsistently by the panel indicated by a significant $F_{Assessor \times Sample}$. Although these attributes would typically be considered as key attributes for evaluating individual assessor performance, the within-group variations render the results unreliable. Only *Milky flavour* and *Cocoa flavour* supposedly emerged as key attributes for which the samples were reliably differentiated. Yet, the individual assessor ANOVAs (**Table 5.5**) showed no significant sample differences ($\alpha=0.05$) for any assessor on these attributes, suggesting that all the assessors may require retraining.

While agreement across assessors ($F_{Assessors}$) may often be overlooked in practice (Næs et al., 2010), it can sometimes influence the identification of key attributes when basing this on overall panel sample discrimination. Significant differences in the Sample effect in the panel ANOVA are based on averaged scores across all the

assessors, which can smooth out inconsistencies in individual ratings. Consequently, even if assessors demonstrate poor repeatability and fail to discriminate between samples individually, pooling their data at the group level can reduce the noise, revealing consistent trends that result in significant differences. Raithatha and Rogers (2018) show how an assessor's poor replication can contribute to non-discrimination of samples especially when the variability in scoring is high for a specific sample. This may have been responsible for the loss of sample discrimination for individual assessors for *Milky flavour* and *Cocoa flavour* in this study.

The Rasch approach determines key attributes more effectively as it estimates each parameter independently based on average ratings across all facets, while accounting for variability in assessors' individual tendencies or biases in scale use. This ensures that the results are not influenced by these variations. The OUTFIT mean square, a residual fit statistic in the Rasch model, identifies responses that deviate from the model's expectations. Although these variations are controlled when estimating sample measures (TIM), assessors whose behaviour deviates from the group are flagged by their OUTFIT Mnsq values, as demonstrated in section **5.3.4**. Additionally, OUTFIT Mnsq for the attributes highlight the contribution of each attribute to product differentiation after accounting for individual differences in scale use, further improving the precision in determining key attributes.

5.3.3.3.1 Key attributes as determined by the Rasch Model

As discussed in the previous chapter (section **4.3.5**), the Rasch model assumes equal discrimination between attributes. In this context, the OUTFIT values for each attribute reflect the variability in its ratings across different samples relative to that of other attributes in the analysis. This provides a clearer indication of the key discriminating attributes. However, it captures unexpected measurement variance arising from two sources: between-group variations (differences in how attributes are rated across samples) and within-group variations (differences caused by interactions between modelled facets, such as assessors or repetition), as previously discussed in section 4.3.4: **pg. 90** (Linacre, 1995).

OUTFIT values below the acceptable range (overfit) indicate that an attribute does not reliably discriminate between samples, whereas values above the acceptable range (underfit) highlight the key attributes driving the differences amongst the samples. Meanwhile as discussed in the previous chapter (section **4.3.5)** , when underfit occurs with a low-intensity attribute, it suggests that, although the attribute may discriminate between samples, its ratings are strongly influenced by individual assessor variations in interpreting the attribute or internal inconsistencies. In some cases, however, the underfit may result from unusually low ratings for a single sample, which pull down the average rating for that attribute across all samples. When this occurs, the underfit for the low-intensity attribute is not mainly due to assessor inconsistencies, but rather due to this downward shift in the mean, which ultimately pulls the affected attribute to the negative side of the logit scale, as demonstrated later with the untrained panel's rating of *Orange flavour* (**Figure 5.5**).

Insights from the attribute outfit statistics can help analysts decide whether an attribute should be removed from the analysis or whether further training is needed for the panel or specific assessors.

Figure 5.3 illustrates the attribute contributions to the overall difference between the chocolate spread samples for the trained panel. As discussed in section **3.3.1.6: Residual fit statistics** , acceptable OUTFIT ranges are context-dependent and can be calculated using the formula***** (Wu & Adams, 2013; Eckes, 2023). For this panel, the acceptable range is between 0.64 and 1.36, based on 63 responses per attribute. and *Milky flavour* emerged as the key discriminating attributes. *Orange flavour* was underfit with an OUTFIT Mnsq values of 1.92, indicating greater variability in ratings across samples, while Milky flavour, with an OUTFIT Mnsq of 0.73, fell well within the acceptable range, showing lesser variations and relatively more consistent ratings across the samples. In contrast, *Sweetness* (0.65) was nearly overfit, and *Cocoa flavour* (0.47) showed clear overfit, indicating that their rating patterns were overly predictable by the model, and thus did not contribute meaningful differentiation between the samples.

---

* $1 \pm 2\sqrt{\frac{2}{Nr}}$, where Nr (number of responses) for each of the attributes is 63 for trained panel and 216 for the untrained panel, respectively.

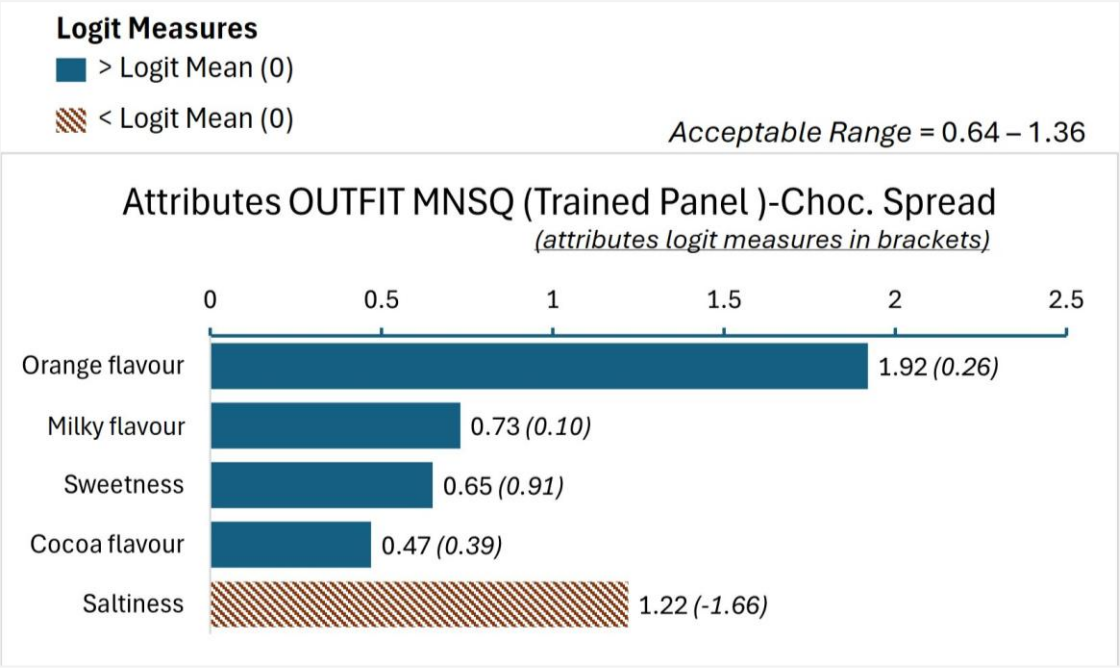


Figure 5.3. Attribute contributions to overall product differences for the Trained Panel based on Rasch logit measures (in brackets) and residual fit statistic (OUTFIT Mnsq). Attributes are colour-coded by logit sign: blue fill= positive logits (higher intensity, contributing more to product differences); red textured fill= negative logits (lower intensity, rated more inconsistently).

Further investigations were made using the Assessor x Sample interaction plots (**Figure 5.4**), trellis plots for each assessor showing the raw data distribution (Section **5.3.4**), and two-way ANOVAs for individual assessors (**Table 5.5**).

The ANOVA results showed that five out of seven assessors identified significant differences between the samples for *Orange flavour* ($p<0.01$), aligning with its high underfit value of 1.92. Minor inconsistencies in the rank order between Brands A and C were also reflected in the same OUTFIT value, even though differences between those samples were not significant as shown in **Table 5.4** below. While the sample effect for *Milky flavour* was not significant for any assessor in the individual ANOVA results (**Table 5.5**), the panel interaction plots in **Figure 5.4** below, and their individual trellis plots in **Figure 5.8** below showed that most assessors consistently rated Brand B as significantly higher in *Milky flavour*, whereas Brand B and the Control were rated similarly.

In contrast, only one assessor detected a slightly significant difference for *Cocoa flavour* ($p<0.10$) according to the individual ANOVA results in **Table 5.5**, while two assessors did for *Sweetness* ($p<0.05$).

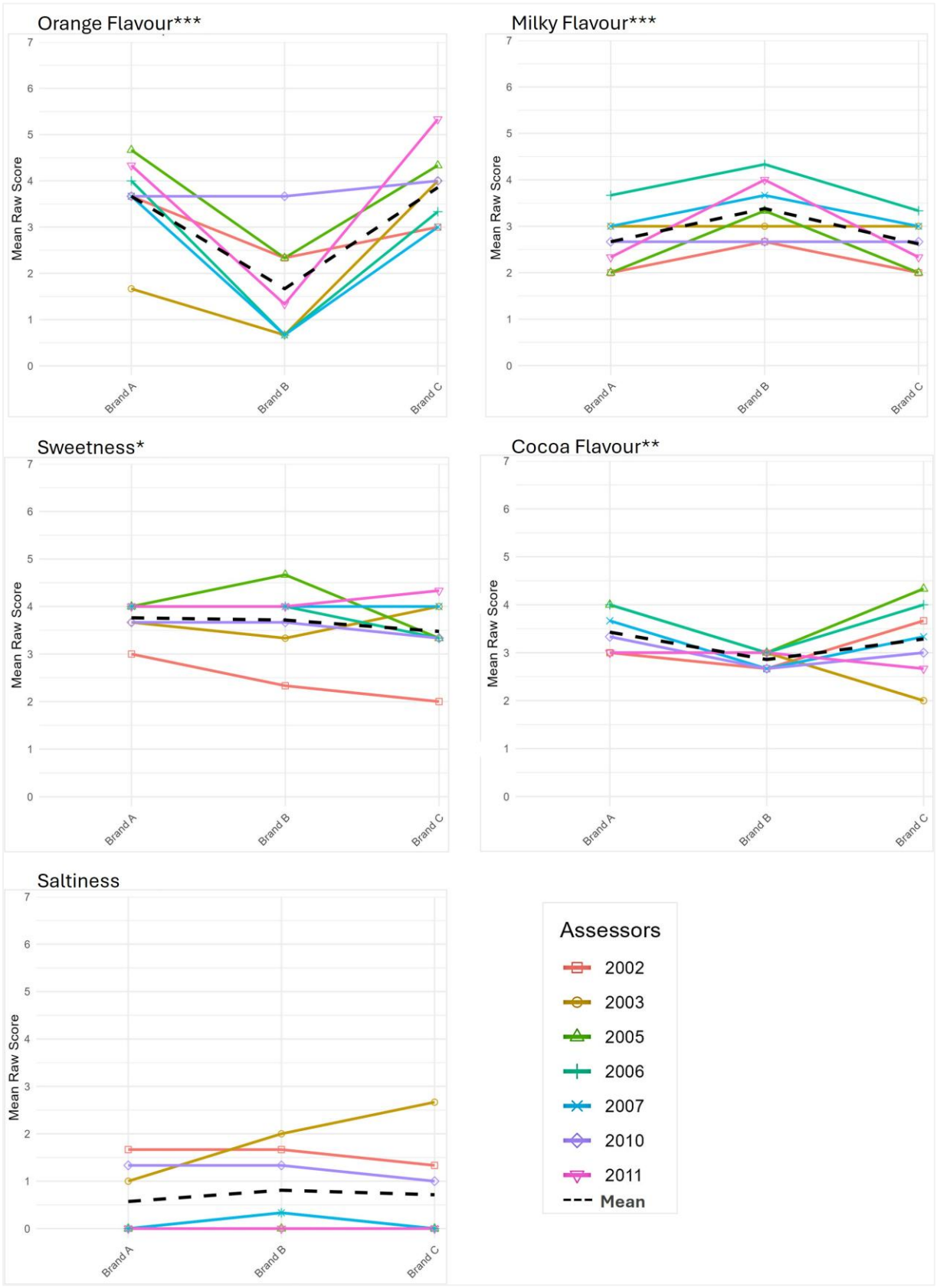


Figure 5.4. Trained panel interaction plots for all attributes.
Attribute titles indicate F-values from panel ANOVA results, with p-value levels significance: <0.001***, <0.01**, <0.05*; measures with no superscript symbols >0.05.

From the interaction plot above, *Sweetness* and *Cocoa flavour* showed both crossover interactions, where the ratings on the samples by some assessors were reversed, and magnitude interactions, where assessors differed in rating severity (Stone et al., 2012). The magnitude effects likely contributed to the loss of discriminatory power for these attributes as indicated by the OUTFIT Mnsq, since the Rasch model already accounts for such interactions when estimating the measures. Crossover interactions, which are reflected in the OUTFIT Mnsq, were more pronounced for *Sweetness* than for *Cocoa flavour*, explaining the slightly higher value observed for *Sweetness*.

Additionally, the trellis plots for the few discriminating assessors revealed poor rating repeatability, possibly due to adaptation or altered sensitivity across replicate evaluations (Sipos et al., 2021). As noted by Stone et al. (2012), crossover interactions reflect insensitivity to the differences between products, unless, perhaps, there are no noticeable differences between the samples. In that case, they may reflect a failure to use the scale correctly, as assessors may be uncertain about whether they are perceiving the attribute and may assign internally inconsistent ratings. In this study, however, there were formulation differences for sweetener and cocoa content (**Table B 2**), so the crossover interactions for *Sweetness* and *Cocoa flavour could* be reflecting that some assessors struggled to detect these differences reliably, resulting in inconsistent ratings.

Saltiness, a low-intensity attribute (logit measure = -1.66), was flagged as problematic due to its underfitting OUTFIT Mnsq value of 1.22. Assessor ratings were inconsistent, showing significant magnitude and crossover interactions, with one assessor revealing a slightly significant difference in *Saltiness* intensity ($p<0.1$) in the individual ANOVA (**Table 5.5**).

In all, the interaction plots revealed crossover and magnitude effects for most attributes, except *Milky flavour*, suggesting inconsistent use of the rating scale and varying sensitivity to product differences among assessors. This is likely due to the lack of specific training on scale use in this study, leading to inconsistencies that undermine both the reliability of the panel and the validity of the results. Further targeted training, particularly on scale familiarisation, is recommended to improve panel alignment.

For the untrained panel, the acceptable OUTFIT Mnsq range was 0.81-1.19 (Nr = 216). Despite the panel's unreliable results, the model still offers valuable insights into the underlying reasons, as shown in **Figure 5.5**. Rating patterns were generally inconsistent, as indicated by the interaction and trellis plots (**Figure 5.6** and **Figure 5.10**, respectively), and these were reflected in the attributes' OUTFIT Mnsq values.

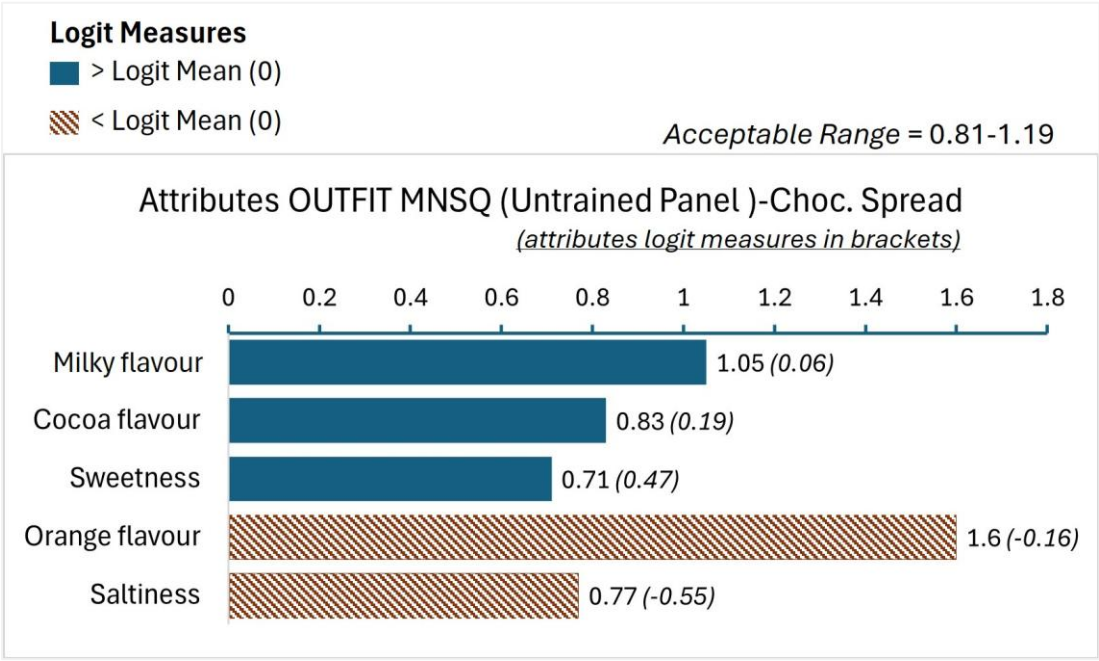


Figure 5.5. Attribute contributions to overall difference for the Untrained Panel based on Rasch logit measures (in brackets) and residual fit statistic (OUTFIT Mnsq). Attributes are colour-coded by logit sign: blue = positive logits (higher intensity, contributing more to product differences); red textured fill= negative logits (lower intensity, rated more inconsistently).

Milky flavour and *Orange flavour* were identified as the key attributes, with corresponding OUTFIT Mnsq values of 1.05 and 1.60, respectively. *Orange flavour* was the strongest contributor to perceived differences among the samples, but Brand B was generally rated as having the lowest intensity for *Orange flavour*, with ratings ranging from barely detectable to non-existent. This is a clear example where an otherwise high intensity attribute received a low rating on one sample, resulting in a low logit measure (-0.16) as previously discussed in section 4.3.5: **pg.96**.

For *Milky flavour*, Brand B was consistently distinguishable from the others by most assessors, and crossover interactions appeared less pronounced compared to those observed with *Orange flavour*, suggesting a more stable perception. *Saltiness* and

Sweetness were flagged as redundant, with OUTFIT Mnsq values of 0.77 and 0.71, respectively, indicating they did not meaningfully differentiate between the samples.

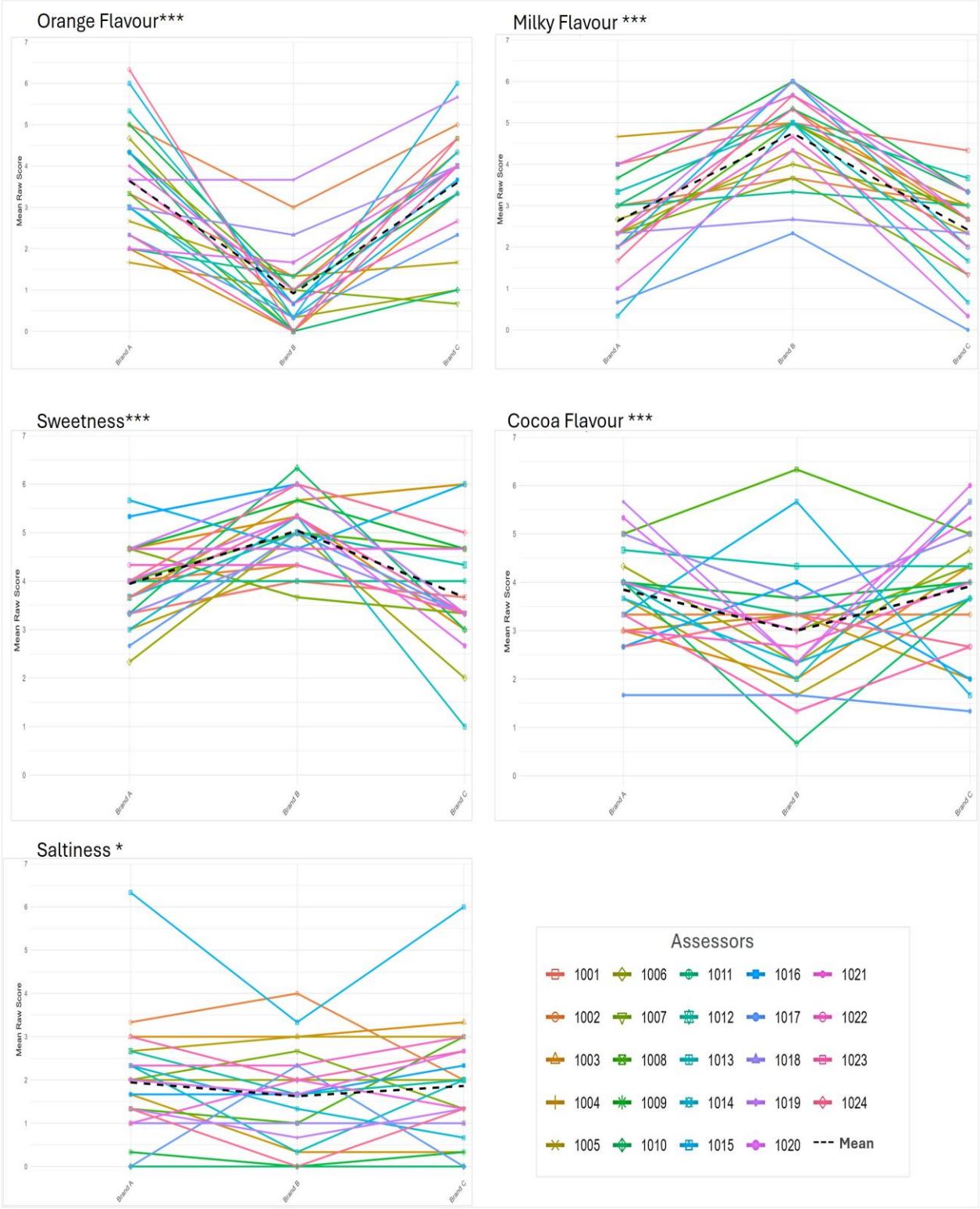


Figure 5.6. Untrained panel interaction plots for all attributes.
Attribute titles indicate F-values from panel ANOVA results, with p-value levels significance: <0.001***, <0.01**, <0.05*; measures with no superscript symbols >0.05.

Cocoa flavour, on the other hand, approached overfitting with a value of 0.83, as the interaction plot revealed a higher degree of crossover interactions, leading to the higher OUTFIT Mnsq value.

Orange flavour and *Milky flavour* were the primary contributors to actual product differences, even for the inconsistent untrained panel. In contrast, *Sweetness* and *Cocoa flavour* were not reliable key attributes for either panel, as the presence of crossover interaction effects, reflected more individual differences and noise than actual product differences.

Saltiness was a redundant attribute for both panels, but one assessor in the trained panel had highly unusual ratings, which caused it to be flagged as a challenging attribute. In the next section, this assessor is also flagged as inconsistent in the assessor OUTFIT analysis.

5.3.3.3.2 Product differences

Table 5.4 presents the sample comparisons based on Kruskal-Wallis mean rank sums of the Rasch measures and Tukey's HSD for the individual attributes.

Table 5.4. Product comparison results for both panels based on Tukey's HSD on raw score mean ratings for individual attributes and Kruskal-Wallis test on mean rank sums of Rasch measures for overall sample comparison.

	Sample[1]	Orange Fl	Milky Fl	Cocoa Fl	Sweetness	Saltiness	*Rasch Measure*[2]
Trained Panel	**Brand A**	3.67^{b}	2.67^{a}	3.43^{b}	3.76^{a}	0.57^{a}	***-0.25***^{b}
	Brand B	1.67^{a}	3.38^{b}	2.86^{a}	3.71^{a}	0.81^{a}	***-0.60***^{a}
	Brand C	3.86^{b}	2.62^{a}	3.29^{ab}	3.48^{a}	0.71^{a}	***-0.28***^{b}
Untrained Panel	**Brand A**	3.64^{b}	2.62^{a}	3.85^{b}	3.94^{a}	1.94^{a}	***-0.28***^{a}
	Brand B	0.92^{a}	4.75^{b}	3.00^{a}	5.04^{b}	1.62^{a}	***-0.33***^{a}
	Brand C	3.60^{b}	2.42^{a}	3.92^{b}	3.67^{a}	1.86^{a}	***-0.32***^{a}

[1] Sample differences based on Tukey's HSD analysis of raw score sample means across individual attributes and Kruskal-Wallis mean rank sums for Rasch measures, where sample Rasch measures with different superscript letters are significantly different ($p<0.05$). Fl=Flavour.

[2] Rasch measures of samples are estimated based on average intensity ratings (Total Intensity Measure -TIM) across all modelled facets (assessors, repetitions, and attributes) after accounting for their influences.

In relation to the product compositions (**Table B 2**), the differentiating ingredients were orange flavouring, milk content, and whether sugar or sugar replacers were used. Based on these differences, it was expected that Brand B would be significantly different from the other samples in terms of *Orange* and *Milky flavour*, as it was not a chocolate-orange spread like the other two and contained higher milk content from both milk chocolate and added milk powder. This expectation was confirmed by the attributes OUTFIT for both panels (**Figure 5.3** and **Figure 5.5**), and the mean raw intensity ratings from the assessors (**Table 5.4**).

While the trained panel rated the *Orange flavour* for Brand B at a higher average intensity than the untrained panel, the extremely low ratings given by most assessors in the untrained panel (see individual trellis plots in **Figure 5.10**) lowered the overall mean intensity across all samples, resulting in a negative logit value for *Orange flavour*, as previously discussed. However, the untrained panel appeared to accurately score the absence of *Orange flavour* in Brand B, with many assessors scoring it as zero. In contrast, the trained panel may have been influenced by expectation error (Meilgaard et al., 2025), anticipating *Orange flavour* in Brand B due to its presence in the other samples, or they may have been playing it safe with their ratings, perhaps a consequence of receiving feedback during training (Myford & Wolfe, 2004; Castura et al., 2005).

Although a significant difference in *Cocoa flavour* might have been anticipated, since Brand B primarily used more milk chocolate crumbs than fat-reduced cocoa mass (**Table B 2**), whereas the other brands used only fat-reduced cocoa mass, the assessors in both panels were unable to reliably distinguish between the samples as discussed earlier (section 5.3.3.3.1, **p131**). Therefore, the reliability and validity of the results regarding the product differences are questionable. Milk chocolate has been shown to be perceived as sweeter and characterised by milk flavour notes, while dark chocolate tends to have more bitter notes (Liu et al., 2015). However, since all samples had similar sugar or sweetener levels (total carbohydrates ~50g), and the cocoa mass in the other brands would have been sweetened as a result, it was hypothesised that any differences in cocoa content due to the addition of milk chocolate would be reflected more as *Milky flavour* and *Sweetness*, rather than as differences in *Cocoa flavour*.

Recall from section **5.3.1** (**Table 5.1**) that the Principal Component Analysis of Residuals (PCAR) indicated possible local dependency between *Sweetness* and *Milky flavour*, as well as between *Cocoa flavour* and *Milky flavour*. This supports the hypothesis that differences in cocoa content from milk chocolate were expressed more strongly through variations in *Milky flavour* and *Sweetness* than in *Cocoa flavour* itself. Just like with *Orange flavour*, the untrained panel appeared to be more sensitive to the product differences. They rated Brand B highest in *Sweetness* and *Milky flavour*, and lowest in *Cocoa flavour*, as shown in **Table 5.4.** In contrast, the trained panel reflected this impact only in *Milky flavour*, with minimal differences observed for *Cocoa flavour*.

Regarding *Sweetness*, Brand C used maltitol as a sugar replacement, while the other two brands used similar amounts of sugar. However, this may not have contributed to noticeable differences, as maltitol, a sugar alcohol, is known to have characteristics very similar to sucrose, except for its lower glycaemic index (O'Donnell, 2012).

Saltiness had the lowest intensity as was reflected on the Wright maps. The salt content for the products ranged from less than 0.01 to 0.13. Low-intensity attributes like this can be difficult to rate accurately. This challenge was further supported by the OUTFIT Mnsq values, which flagged *Saltiness* as problematic for the trained panel, and non-discriminating for the untrained panel. The lack of discrimination in the untrained panel was most likely due to inconsistent ratings and poor replication, whereas for the trained panel, the issue was traced to a single disagreeing assessor.

Myford and Wolfe (2004) have demonstrated that the performance of individual raters in the Rasch model is assessed relative to the group being evaluated, and that deviations from model expectations, such as those indicated by OUTFIT Mnsq also depend on this context. In their study, this meant that an individual rater's fit statistic would highlight when their ratings did not align with the rest of the panel. In the present study, this principle explains why the trained panel's overall consistency made one assessor's disagreement stand out clearly, whereas in the untrained panel, inconsistent scoring across multiple assessors resulted in overfit, since no single rater's pattern stood out enough to affect the OUTFIT Mnsq (discussed further in section **5.3.4**). This demonstrates how interpretations of

assessor performance and attribute discrimination depend on the panel's collective pattern of responses. In this sense, the approach mirrors the conventional method, where individual assessor performance is evaluated relative to the panel mean, which serves as the reference standard in the absence of a known true attribute mean (Stone et al., 2012; Raithatha & Rogers, 2018; British Standards Institution, 2021; Meilgaard et al., 2025). However, from the outset, the Rasch model adopts a diagnostic perspective, encouraging a deeper investigation of the data through diagnostic tools that are integrated within a single analysis. The Rasch-approach to examining individual assessor performance is explored in the following section.

5.3.4 Comparison of individual assessor performance for both panels

The performance of individual assessors was evaluated based on their discriminatory ability, internal consistency or repeatability, relative consistency with other assessors, and rating effects or rater bias. Two-way ANOVAs and the Rasch model's quality control statistics were used to analyse the data.

Discriminatory ability refers to the proportion of key attributes on which an assessor can distinguish the samples. Internal consistency refers to an assessor's ability to consistently rate samples across replicate evaluations and is also referred to as repeatability. Relative consistency refers to an assessor's ability to assign similar intensity ratings and rank samples the same way as other assessors in a panel, termed "agreement across assessors" by ISO 1132:2021 (British Standards Institution, 2021). Rating effects refers to differences in scale usage by each assessor.

In practice, several techniques are employed to examine panel performance criteria. Bárcenas et al. (2000) demonstrated that conducting ANOVA-based comparisons of assessor F-ratios and residuals can effectively identify assessors who contribute most to variability, allowing detection of inconsistent scoring behaviour. Tomic et al. (2007) further showed that visualising these results using graphical methods such as eggshell or correlation plots provides a clearer overview of assessor differences and panel agreement, and that the most comprehensive understanding of panel performance is achieved by combining analytical statistics

with several visualisation techniques. This recommendation was echoed by Stone et al. (2012), Ho (2015) and Raithatha and Rogers (2018).

More recent guidelines continue to rely on these established diagnostic techniques, using them in combination with mixed ANOVA models, correlation analyses between assessor scores and the panel mean, and advanced graphical tools (British Standards Institution, 2021; Meilgaard et al., 2025). While these methods are comprehensive, they can be demanding for the panel leader, especially when a rapid overview is required to identify assessors who may be deviating from the group.

Using the ANOVA approach, individual two-way ANOVAs are conducted for each assessor across all attributes. The Sample main effect (F_{Sample}) indicates discriminatory ability. Agreement across assessors is captured in the three-way ANOVA for the panel discussed earlier (**Table 5.2**) and is inversely related to the Assessor main effect ($F_{Assessor}$) (ISO 1132:2021 (British Standards Institution, 2021)). However, poor scoring repeatability can mask this relationship.

Each assessor's internal consistency in assigning ratings across replicate evaluations, is estimated using the interaction effect between sample and the replicate session/repetition ($F_{Sample \times Repetition}$) for each assessor, derived using the Tukey's additivity test[1] (Ho, 2015). A significant interaction term indicates inconsistency. While this provides valuable information, it does not offer a complete picture. Better insights can be gained by combining the analyses with response distribution plots(Stone et al., 2012; Ho, 2015) and applying one or more of the complementary methods described above.

The ANOVA approach assesses discrimination, agreement, and repeatability, but does not fully capture assessor severity, scale usage, or consistent understanding of attributes. These factors can affect data quality but may not be evident from F-ratios alone. Combining statistical results with graphical tools and other diagnostics therefore gives a more complete evaluation and supports targeted panel training.

[1] Tukey's additivity test is used to check whether the interaction between two factors in a two-way ANOVA without replication is negligible, thereby testing if the model assumption of additivity holds true (i.e., that the effects of the factors are purely additive). A significant p-value (typically $p < 0.05$) means the assumption is violated and the interaction term is significant.

Rasch analysis enhances efficiency by providing an overview of disagreeing assessors using inherent quality control statistics. The model accounts for individual rater bias resulting from idiosyncratic use of scale, while estimating Total Intensity Measures (TIM) for the samples (as discussed in section **3.1.2**). It then assigns severity logit measures to each assessor based on their tendency to consistently assign higher or lower ratings relative to the panel, as shown in the Wright maps. All other imprecisions in the response data, including interaction effects, are recorded as outfit (Linacre, 1995) as discussed previously in section **4.3.4**. In their study about measuring rating effects with the Many-Facet Rasch Model, Myford and Wolfe (2004), provide a detailed account on how the model offers insights into various individual rater biases and inconsistencies.

Using Rasch model fit indices for the *Assessor facet*, OUTFIT Mnsq ranges indicate different types of rater bias. As previously discussed, acceptable OUTFIT Mnsq ranges are context-dependent and derived using the formula suggested by* (Eckes, 2023), based on the total number of responses used to estimate the facet parameters. The higher the degree of outfit above 1, the more deviation from the model expectations. Conversely, the lower the OUTFIT Mnsq below 1, the more predictable the response is by the model.

According to Myford and Wolfe (2003) and (2004), when an assessor's OUTFIT Mnsq is overfit, it indicates a lack of variation in their ratings across samples or attributes. This is often due to restriction of range (clustered ratings within a specific portion of the scale) or central tendency bias (overusing middle categories) and reflects the assessor's inability to discriminate between samples. In contrast, when the OUTFIT Mnsq is above the acceptable range (underfit), it suggests that the assessor is either inconsistent in their ratings across replicates or in disagreement with the rest of the panel. Assessors' fit indices indicate their cumulative agreement between observed and expected ratings across all attributes, samples and replicate evaluations. Therefore, the OUTFIT Mnsq reflects both internal and relative inconsistencies. (Myford & Wolfe, 2004) term these inconsistencies the "Randomness effect", where haphazard or seemingly random ratings suggest that the assessor does not reliably

* $1 \pm 2\sqrt{\frac{2}{Nr}}$, where Nr (number of responses) for the assessors in both panels is 45 as each assessor rated the 3 samples, across the 5 attributes in 3 replicates.

differentiate between samples. The model further identifies assessors who rank the samples differently than the rest of the panel using point-biserial measure correlations for the assessor facet, which the authors termed the “Single Rater – Rest of Raters (SR/ROR)” correlation. A lower correlation value in this measure flags assessors whose rankings deviate from the panel’s overall pattern. The assessors’ OUTFIT Mnsq and SR/ROR correlation results will be presented in the following section.

A caveat of the Rasch approach to evaluating individual assessor performance is that the results are always relative to the performance of all other assessors in the analysis. Therefore, it is most informative when applied to a more homogeneous panel rather than to an inconsistent one. The results for the performance of assessors in the trained and untrained panels are presented in the following pages.

5.3.4.1 Performance of trained panel individual assessors

Figure 5.7 presents a control plot for the OUTFIT Mnsq values of the trained panel. The acceptable fit range for assessors was between 0.57 and 1.42. Ratings from assessors within this range were generally consistent with the rest of the panel, whereas those with values greater than 1.42 exhibited signs of random or inconsistent rating behaviour. Assessors 2003 was flagged as slightly inconsistent, with an OUTFIT Mnsq value of 1.42, right at the acceptable threshold compared to the rest of the panel.

The ANOVA results in **Table 5.5**, which also include Rasch model rater performance indices, show that Assessor 2003 was able to distinguish between samples for *Orange flavour* ($p<0.05$) and, to a lesser extent, *Saltiness* ($p<0.10$), while no significant differences were found for the other attributes. Meanwhile, the Rasch model single rater-rest of rater (SR/ROR) correlation provided complementary information, showing that Assessor 2003 ranked the samples differently from the rest of the panel for these attributes, with a correlation value of 0.46, the lowest among all assessors.

However, it should be noted that the Rasch model evaluates assessor fit relative to the collective pattern of responses within the panel. Therefore, although Assessor 2003’s OUTFIT Mnsq value was at the acceptable threshold, the raw data indicate

broader variability across the panel. This suggests that the apparent fit reflects relative consistency with a panel that itself exhibited some instability, rather than absolute rating consistency.

The trellis plots in **Figure 5.8** further support these findings, showing that Assessor 2003 perceived a larger difference in the intensity of *Orange flavour* between the samples compared to the rest of the panel. In addition, they reversed the order in which Brands A and C were rated, assigning a higher intensity rating to Brand C opposite to the pattern observed in most other assessors. They also perceived greater differences in *Saltiness* across the samples compared to the panel average.

SR/ROR correlation values (**Table 5.5**) for Assessors 2002 and 2011 were also relatively low, at 0.51 and 0.77, respectively. While the OUTFIT Mnsq value for Assessor 2002 (1.11) indicated consistency in their ratings, their ANOVA results showed they could not discriminate between the samples based on *Orange flavour*, a key attribute. However, they could differentiate *Sweetness* ($p < 0.05$), which was not a key attribute. Additionally, the trellis plot indicated that, although their *Saltiness* ratings were somewhat erratic, they observed a large difference in *Saltiness* in at least one replicate evaluation.

Assessor 2011, with an OUTFIT Mnsq value of 1.41, was again close to the misfit threshold (1.42). This was likely due to a reversal in the order of *Orange flavour* intensity between Brands A and C, like Assessor 2002. However, the difference between the samples was not significant, consistent with the rest of the panel, suggesting that the crossover interaction likely resulted from uncertainty about which sample had the higher intensity. Further training could help increase their sensitivity and improve the refinement of their ratings.

Assessor 2010, on the other hand, exhibited overfitting, with an OUTFIT Mnsq value of 0.44, suggesting they were using a restricted range of the scale and likely not discriminating between the samples. This was confirmed by their ANOVA results, where none of the attributes showed significant differences, and the trellis plot revealed that their ratings never exceeded two scale categories across all attributes.

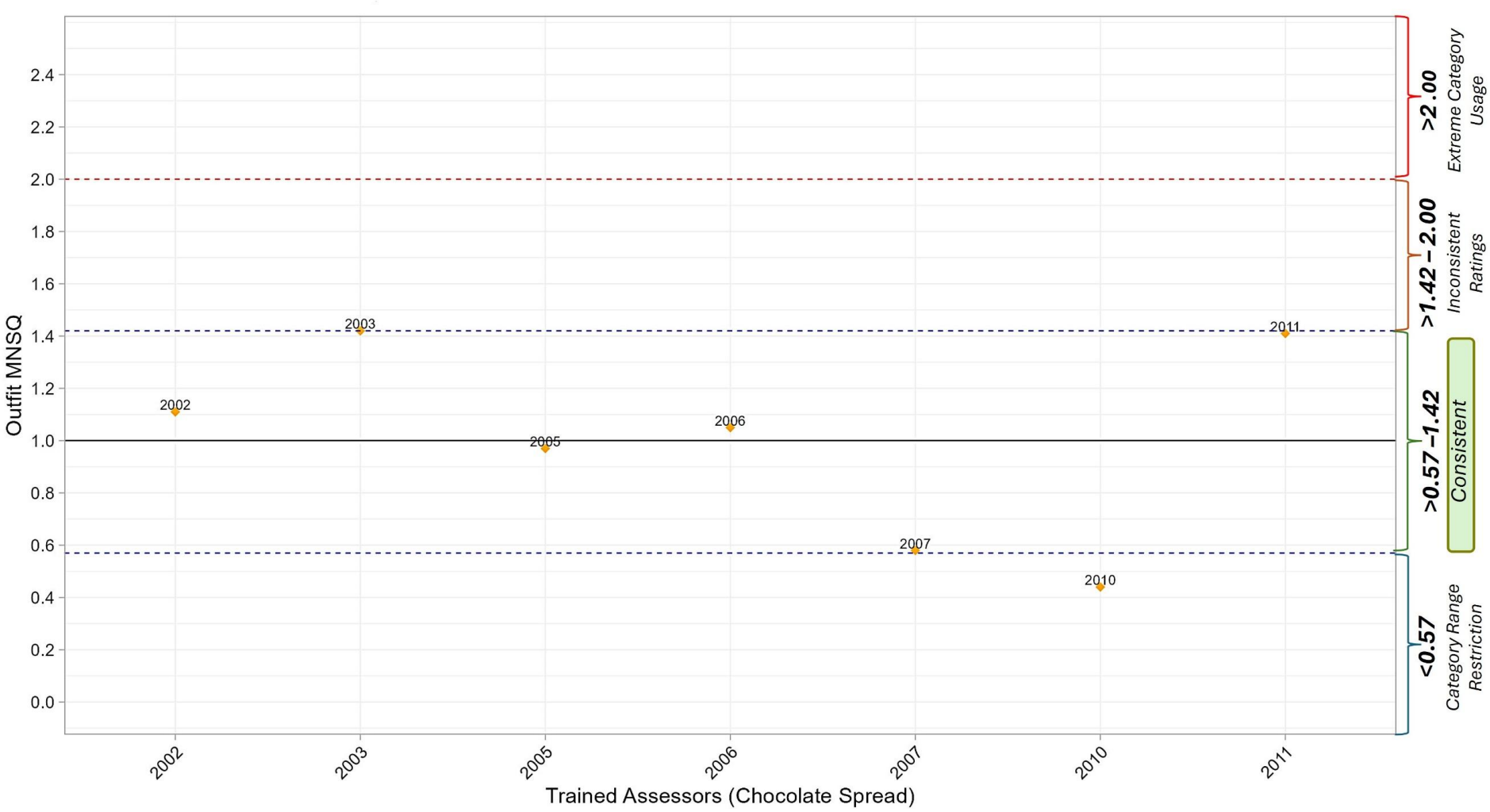


Figure 5.7. OUTFIT Mnsq plot for assessors in the Trained panel.

Table 5.5. Summary of individual ANOVA results for the trained panel, showing Rasch model indicators for rater performance.

Assessor	Rasch Model Indices		Orange Flavour***			Milky Flavour***			Sweetness*			Cocoa Flavour**			Saltiness		
	OUTFIT[1]	SR/ROR[2]	F_{SA}[3]	F_{Rep}	$F_{SA\ X\ Rep}$	F_{SA}	F_{Rep}	$F_{SA\ X\ Rep}$	F_{SA}	F_{Rep}	$F_{SA\ X\ Rep}$	F_{SA}	F_{Rep}	$F_{SA\ X\ Rep}$	F_{SA}	F_{Rep}	$F_{SA\ X\ Rep}$
2002	1.11	0.51	2	3.5	1.5	4	1	NA[4]	7*	1	0.1	2.8	0.4	0.4	0.1	0.6	0.1
2003	1.42	0.46	7.9*	0.4	0.3	NA	NA	NA	2	2	1.8	NA	NA	NA	4.8~	1	0.3
2005	0.97	0.86	43**	7*	0	4	1	NA	8*	2	0	5.2~	2.8	2	NA	NA	NA
2006	1.05	0.81	28**	9*	2.3	2.8	1.6	0	1.6	2.8	0	3	1	NA	1	1	NA
2007	0.58	0.88	26.8**	0.4	0	4	1	NA	NA	NA	NA	2.8	1.6	0	1	1	NA
2010	0.44	0.81	0.25	1.8	2.3	0	0	NA	0.4	1.6	0.3	1	0	NA	1	4	NA
2011	1.41	0.77	13*	1	0.9	2.9	1.9	8.3~	1	19**	5.4	0.3	1	0.1	NA	NA	NA

[1] OUTFIT mean square range: 0.57-1.42 (Nr = 45). Values <0.57 indicate overfit, >1.42 underfit, and >2.0 suggest use of extreme categories. Row shading reflects OUTFIT interpretation: **blue** (overfit) for restriction of range and brown (underfit) for relative inconsistency.

[2] Single Rater–Rest of Rater (SR/ROR) correlation, where values noticeably lower than those of other assessors indicate that an assessor is ranking samples in a different order from the panel. Grey-shaded cells highlight assessor response patterns that deviate from the panel.

[3] F-values with p-value levels of significance: <0.001***, <0.01**, <0.05*; <0.10~ measures with no superscript symbols >0.10. Where SA = Sample, Rep=Repetition and SA X Rep= the Sample and Replicate interaction factor. The levels of significance also apply to the list of attributes in the first row showing differences for F_{Sample} in the panel ANOVA (**Table 5.2.**).

[4] NA signifies no variation in assessor ratings, limiting the ANOVA model's ability to estimate the contribution of the effect.

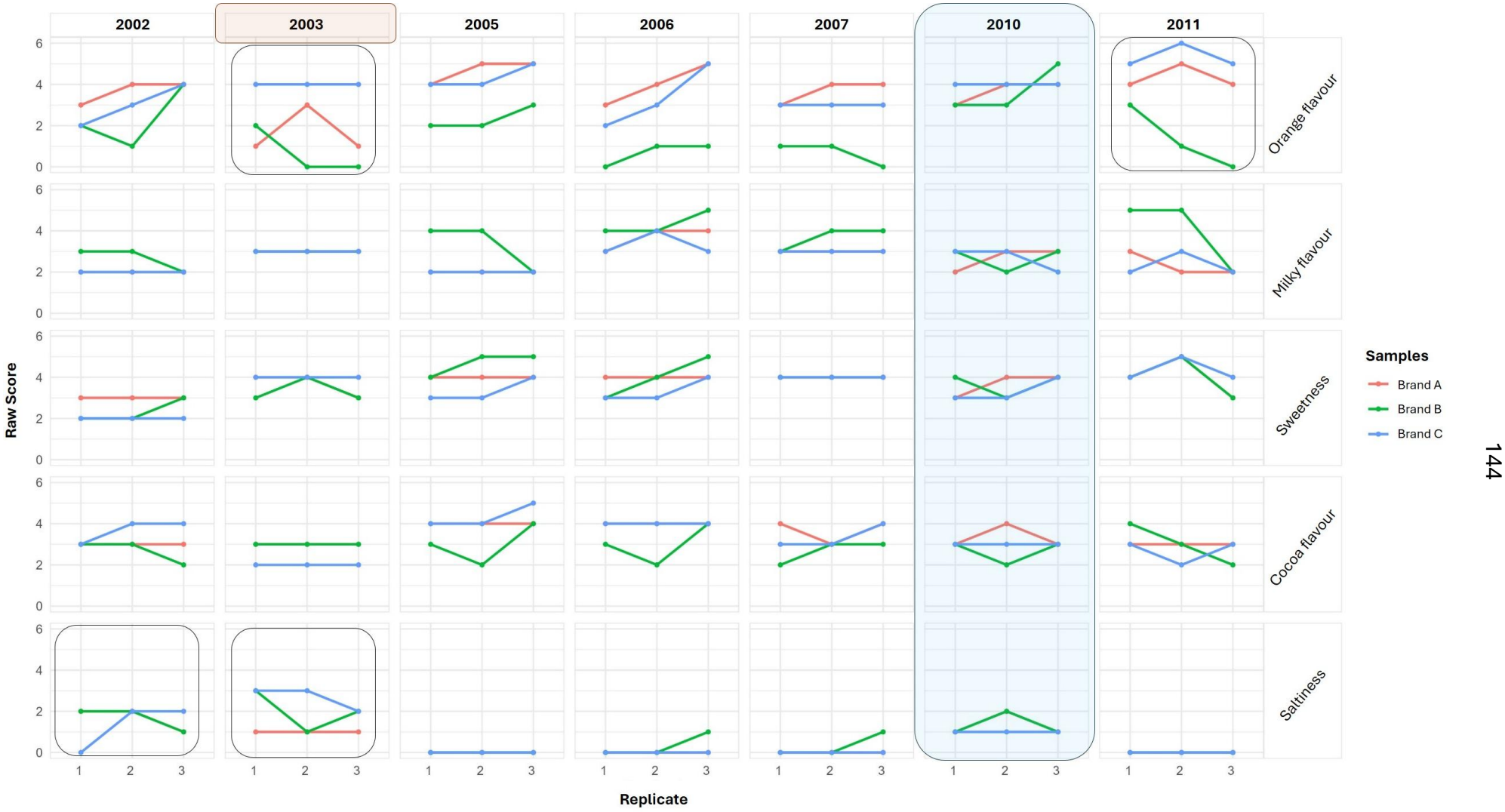


Figure 5.8. Trellis plots for the Trained panel showing the response distribution of raw scores and highlighting model misfit for individual assessors. Shading indicates types of rating effects: **blue** for restriction of range and **brown** for inconsistent ratings. **Grey borders** correspond to assessor response patterns that deviate from the panel, as indicated by low SR/ROR correlation values.

This pattern suggests that something else might be affecting their ratings, perhaps they were distracted. Further investigation into the "any other comments" section of the questionnaire revealed that Assessor 2010 was the only one to leave a comment, which simply stated "Bitter" for 8 out of the 9 evaluations, suggesting some level of disengagement.

The performance of the trained panel was more inconsistent than expected. While a few assessors demonstrated reasonable discrimination, others showed variable or restricted scale use that limited the panel's overall reliability. These findings suggest that targeted retraining and scale calibration would be required to improve both individual performance and collective panel agreement. The Rasch model supported these observations by identifying assessors whose rating behaviour deviated from the collective panel pattern and by quantifying the extent of misfit and their interpretation. However, the results also showed that acceptable fit values did not always correspond to high data quality, highlighting the importance of examining Rasch outputs in conjunction with raw data visualisations to obtain a complete understanding of panel performance.

5.3.4.2 Performance of untrained panel individual assessors

The control plot for the OUTFIT Mnsq values of the untrained panel is presented in **Figure 5.9,** using the same acceptable fit range of 0.57 to 1.42 as applied to the trained panel, since they provided the same number of responses. Unlike the trained panel, assessors in this group exhibited more erratic rating patterns, poor repeatability, and used extreme scale categories, as seen in the trellis plots in **Figure 5.10**. Assessors 1005, 1007, 1011, 1012, and 1018 were flagged as overfit, with OUTFIT Mnsq values of 0.56, 0.56, 0.15, 0.51, and 0.38, respectively, suggesting they used restricted parts of the rating scale. In contrast, assessors 1010, 1013, and 1017 were identified as underfit, with OUTFIT Mnsq values of 1.42, 1.46, and 1.48, indicating relatively inconsistent ratings compared to the rest of the panel. Assessor 1015, identified as the most lenient on the Wright map (**Figure 5.2)**, frequently used both the upper and lower extremes of the scale, had the highest OUTFIT Mnsq value of 2.49. These flagged rating behaviours from the Rasch analysis corresponded with patterns observed in the raw score distribution trellis plots (**Figure 5.10**).

Table 5.6 summarises both the Rasch model assessor performance indices and individual ANOVA results. The ANOVA indicated that 14, 11, 9, 9, and 6 assessors could detect significant differences ($\alpha = 0.05$) between the samples based on *Milky flavour*, *Orange flavour*, *Sweetness*, *Cocoa flavour*, and *Saltiness*, respectively. At a less stringent level ($\alpha = 0.10$), a few additional assessors identified differences: 6 for *Orange flavour*, 4 for *Milky flavour*, 2 for *Sweetness*, and 1 for *Cocoa flavour*. These assessors likely contributed to the highly significant panel discrimination ($p < 0.001$) across all attributes, with *Saltiness* being significant to a lesser extent ($p < 0.05$). However, interaction plots revealed substantial crossover and magnitude interaction effects, while trellis plots showed poor replication across most attributes. Only *Orange* and *Milky flavours* exhibited more consistent patterns.

Assessors flagged as overfit are shaded blue in **Figure 5.10**, with black borders indicating central tendency effects. Lower OUTFIT Mnsq values within the overfit range appeared to indicate a central tendency effect, where ratings were restricted to the middle categories across attributes, as seen in Assessors 1011 and 1018. In contrast, higher OUTFIT Mnsq values within the overfit range pointed toward category range restriction, as observed with Assessors 1005 and 1007. However, it remains unclear whether there is a specific OUTFIT Mnsq range that consistently indicates these effects.

The Rasch model's rater performance indices are relative measures, indicating an assessor's performance compared to the rest of the group (Myford & Wolfe, 2004). Since many assessors in the untrained panel were inconsistent, the OUTFIT Mnsq and SR/ROR correlation values were less informative for analysts attempting to identify assessors in disagreement, occasionally flagging assessors who had slightly better internal consistency. This was observed for the underfitting Assessor 1013, who generally ranked the samples consistently across replicates, despite some issues with repeatability (**Figure 5.10**). *Saltiness* and *Orange flavour* were exceptions, with replicate evaluations showing crossover interactions. This assessor was able to distinguish between samples for all attributes ($p < 0.001$) except *Saltiness*. *Orange flavour* was only significant at the 10% level ($\alpha = 0.10$), likely due to greater internal inconsistency in both *Saltiness* and *Orange flavour*.

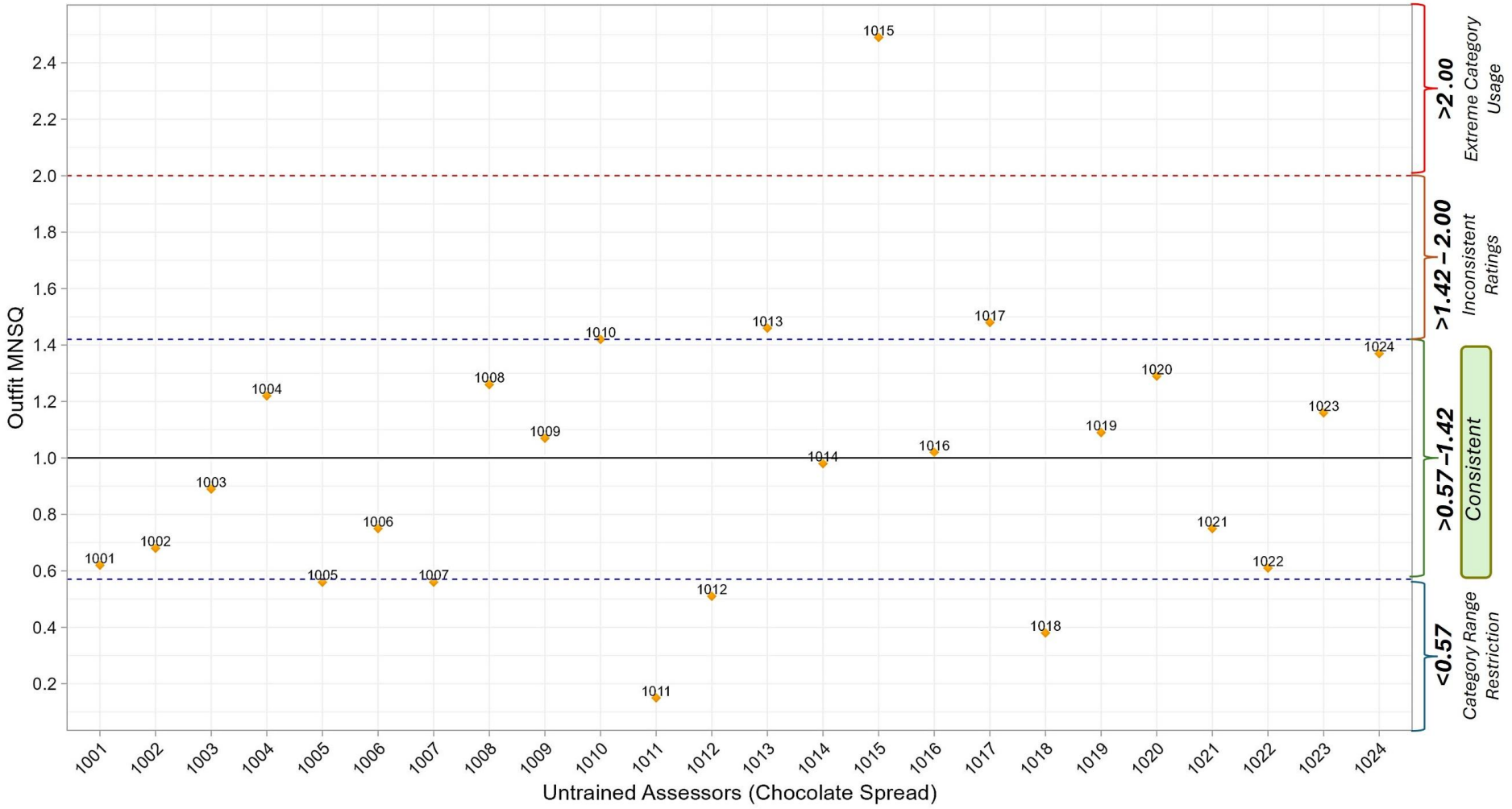


Figure 5.9. OUTFIT Mnsq plot for assessors in the Untrained panel.

Table 5.6. Summary of individual ANOVA results for the Untrained panel, showing Rasch model indicators for rater performance

Assessor	Rasch Model Indices		Orange Flavour***			Milky Flavour***			Sweetness***			Cocoa Flavour***			Saltiness*		
	OUTFIT[1]	SR/ROR[2]	F_{SA}[3]	F_{Rep}	$F_{SA \times Rep}$	F_{SA}	F_{Rep}	$F_{SA \times Rep}$	F_{SA}	F_{Rep}	$F_{SA \times Rep}$	F_{SA}	F_{Rep}	$F_{SA \times Rep}$	F_{SA}	F_{Rep}	$F_{SA \times Rep}$
1001	0.62	0.65	4.3~	1.6	1	7*	1	0.1	2	2	1.8	6.4~	2.8	0	NA[4]	NA	NA
1002	0.68	0.13	3	0.3	1	1	3.3	1.3	2.8	0.4	0	0	1.8	NA	7*	1	0.1
1003	0.89	0.31	3	0.6	0	38**	2	0.3	26**	2	2.6	14.8*	2.8	0.1	0.2	0.2	1.2
1004	1.22	0.64	1.5	0.2	0.1	1.5	1.2	23.7*	43**	4	1.8	0.9	2.6	0.2	16*	7*	0.1
1005	0.56	0.28	1.5	0.1	0.5	12*	1	0.2	5.2	0.4	0.2	4	1	3.9	1	1	NA
1006	0.75	0.39	24.5*	0.5	6.7~	3.3	1	0.5	73***	1	0.6	43**	7*	0.4	NA	NA	NA
1007	0.56	0.55	18.5**	2	0.1	37**	7*	0.1	3.3	0.3	1.3	13*	1	2.8	8*	8*	0
1008	1.26	0.54	7.4*	1.3	0.8	9.6*	1	0	7*	73**	0	4	3.3	0.5	12.4*	6.4~	0
1009	1.07	0.74	6.8~	1.6	0	5.4~	2	1.9	3	13*	49**	0.3	3.3	0.5	1	4	NA
1010	1.42	0.64	6.8~	0.3	3.4	6.1~	4	1.1	13*	0.1	0	3.3	0.1	0.1	NA	NA	NA
1011	0.15	0.78	52**	1	2.8	1	1	NA	NA	NA	NA	4	1	NA	1	1	NA
1012	0.51	0.63	39**	4	49**	14*	2	2.8	4	3	NA	0.3	0.3	0.1	2.8	0.4	0
1013	1.46	0.35	6.9~	1.1	0.3	50**	26**	0.1	19**	1	0	36.4**	2.8	1.4	3.7	0.8	0.6
1014	0.98	0.46	7.3*	1.3	16.8*	30.5**	0.5	0.1	6.1~	2.7	3.2	2.3	0.6	0.1	2.7	0.6	6.3~
1015	2.49	-0.03	28.9*	1.9	0.2	8.7*	0.2	0.1	3.3	1.8	0.1	19.8***	1.3	0.1	10.4*	2.7	1.3
1016	1.02	0.51	10.3*	1.6	0.1	6.5~	3.5	2.1	52**	13*	0.9	2.4	3.8	0.3	0.2	0.4	0.4
1017	1.48	0.43	0.8	0.5	1.5	3.7	2	85.5***	3.7	0.3	1.5	0	1.2	1.2	3.8	1	NA
1018	0.38	0.73	7.6*	2.8	1	0.1	0.3	NA	16*	7*	0.1	16*	1	NA	NA	NA	NA
1019	1.09	0.58	3	4	1.9	37**	0	NA	32**	2	1.8	25**	0.3	0.1	1.6	1.6	0.3
1020	1.29	0.38	0.6	0	1.6	124***	4	0.4	4.6~	0.3	2.8	14.7*	0.1	2	7*	4	0.1
1021	0.75	0.65	4.9~	2.4	4.3	13*	3	0.1	0	7*	NA	3.7	1.9	0.5	2.8	2.8	13.3*
1022	0.61	0.39	5.3~	0.3	0.5	14*	0.5	0.8	3	1	0.2	3.3	1	0.5	1	1.8	2.3
1023	1.16	0.28	12.8*	3.7	15.1*	5.2~	0.4	0.3	3.1	0.7	1.8	1.6	1.1	2	0.3	0.2	0.7
1024	1.37	0.58	48.5**	2	2.6	39**	4	0.1	6	2	0	1	0.3	0.1	1.2	1.2	1.3

[1] OUTFIT mean square range: 0.57-1.42 (Nr = 45). Values <0.57 indicate overfit, >1.42 underfit, and >2.0 suggest use of extreme categories. Row shading reflects OUTFIT interpretation: blue (overfit) = range restriction and non-discrimination (black borders), brown (underfit) = relative or internal inconsistency, and red= use of extreme categories.

[2] Single Rater–Rest of Rater (SR/ROR) correlation, where values noticeably lower than those of other assessors would indicate that an assessor is ranking samples in a different order from the panel. For this panel, no clear trend was observed, as their ratings were generally inconsistent.

[3] F-values with p-value levels of significance: <0.001***, <0.01**, <0.05*; <0.10~ measures with no superscript symbols >0.10. Also applies to the attributes (F_{Sample}) with significant differences in **Table 5.2.**

[4] NA signifies no variation in assessor ratings, limiting the ANOVA model's ability to estimate the contribution of the effect.

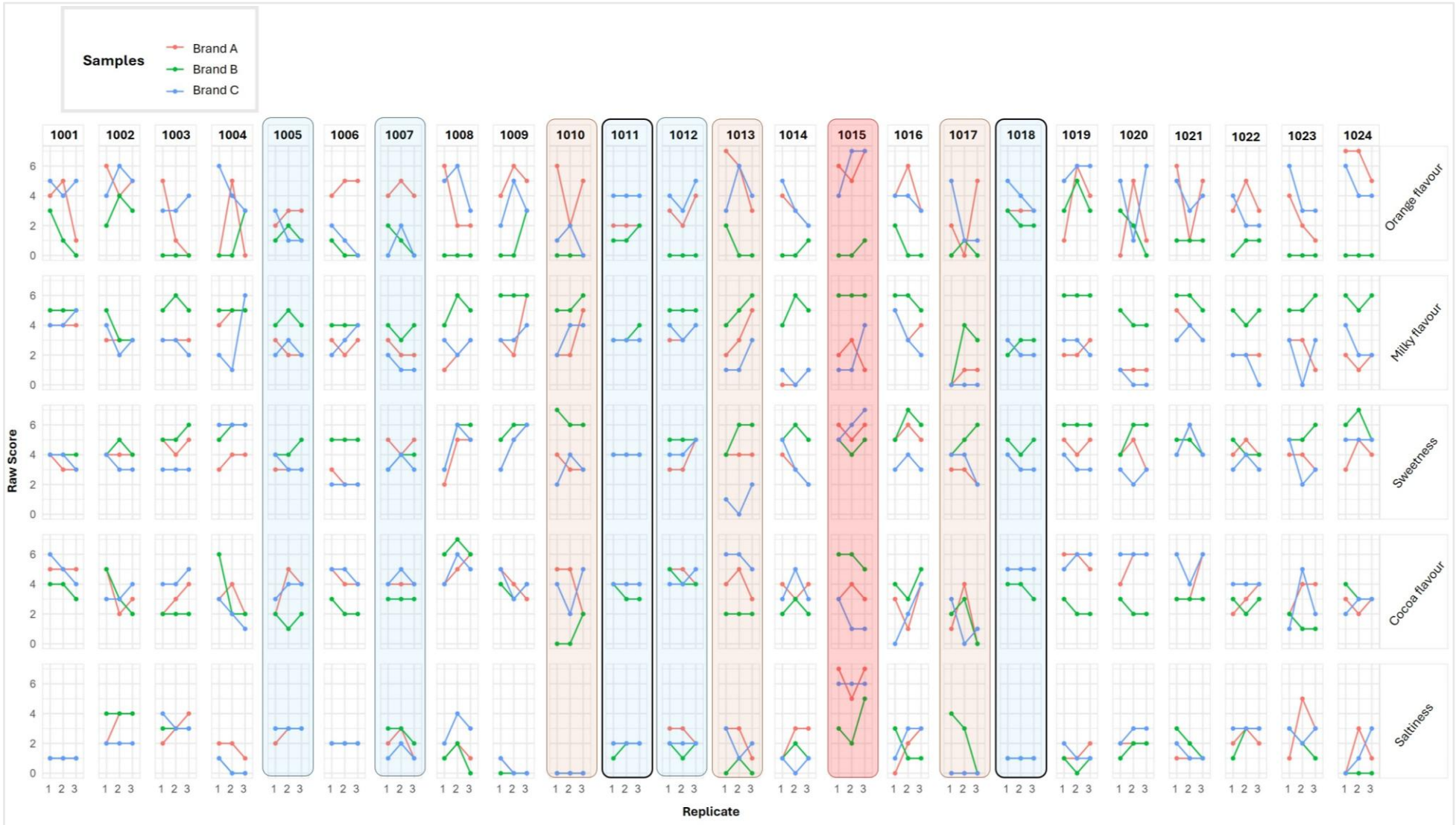


Figure 5.10. Trellis plots for the Untrained panel showing the response distribution of raw scores and highlighting model misfit for individual assessors. Where shading indicates a type of rating effect: **red** for extreme category use; **blue** for restriction of range; **blue with black borders** for central tendency; and **brown**, for inconsistent ratings relative to the expected panel performance.

Other underfitting assessors with trellis plots highlighted in brown (1010 and 1017) seemed to have a higher number on attributes with poor repeatability. Assessor 1017 identified as the most severe assessor on the Wright map (**Figure 5.2**), had the next highest OUTFIT Mnsq value (1.48) after Assessor 1015, who showed extreme misfit.

SR/ROR correlation values were inconsistent across all assessors, offering no clear pattern and limiting their usefulness as a reliable performance indicator for this panel.

Insights from the Rasch performance indices, combined with the ANOVA discrimination results for the key attributes *Orange flavour* and *Milky flavour*, which were identified from the attributes facet outfit analysis, were used to identify assessors whose performance more closely aligned with that of the trained panel for further analysis.

5.3.4.3 Rasch analysis of selected untrained assessors

Eight assessors: 1001, 1008, 1009, 1014, 1016, 1022, 1023, and 1024 were selected from the untrained panel based on their relatively consistent rating patterns, as indicated by their OUTFIT Mnsq values, and their ability to discriminate between samples **(Table 5.6**) using the key attributes, *Orange flavour* and *Milky flavour*, identified by the MFRM. The main aim of selecting these assessors was to determine whether an untrained but carefully screened subgroup could achieve results comparable to those of a trained panel. Their performance was then compared with that of the trained panel using Rasch separation statistics and rater performance indices to evaluate whether similar sample results could be obtained.

The global model fit of the data was acceptable, as only 1% of absolute standardised residuals exceeded 2, suggesting no major inconsistencies that could distort the measurement. The unexplained variance in the first contrast had an eigenvalue of 2.18, slightly above that of the original panel (1.96), indicating a strength in 2 out of 5 attributes and suggesting the possibility of a secondary dimension. A correlation of 0.46 was observed between the standardised residuals of *Sweetness* and *Milky flavour*, confirming that the rating for *Sweetness* was associated with that of *Milky*

flavour, most likely reflecting the impact of milk chocolate in Brand B, as discussed earlier.

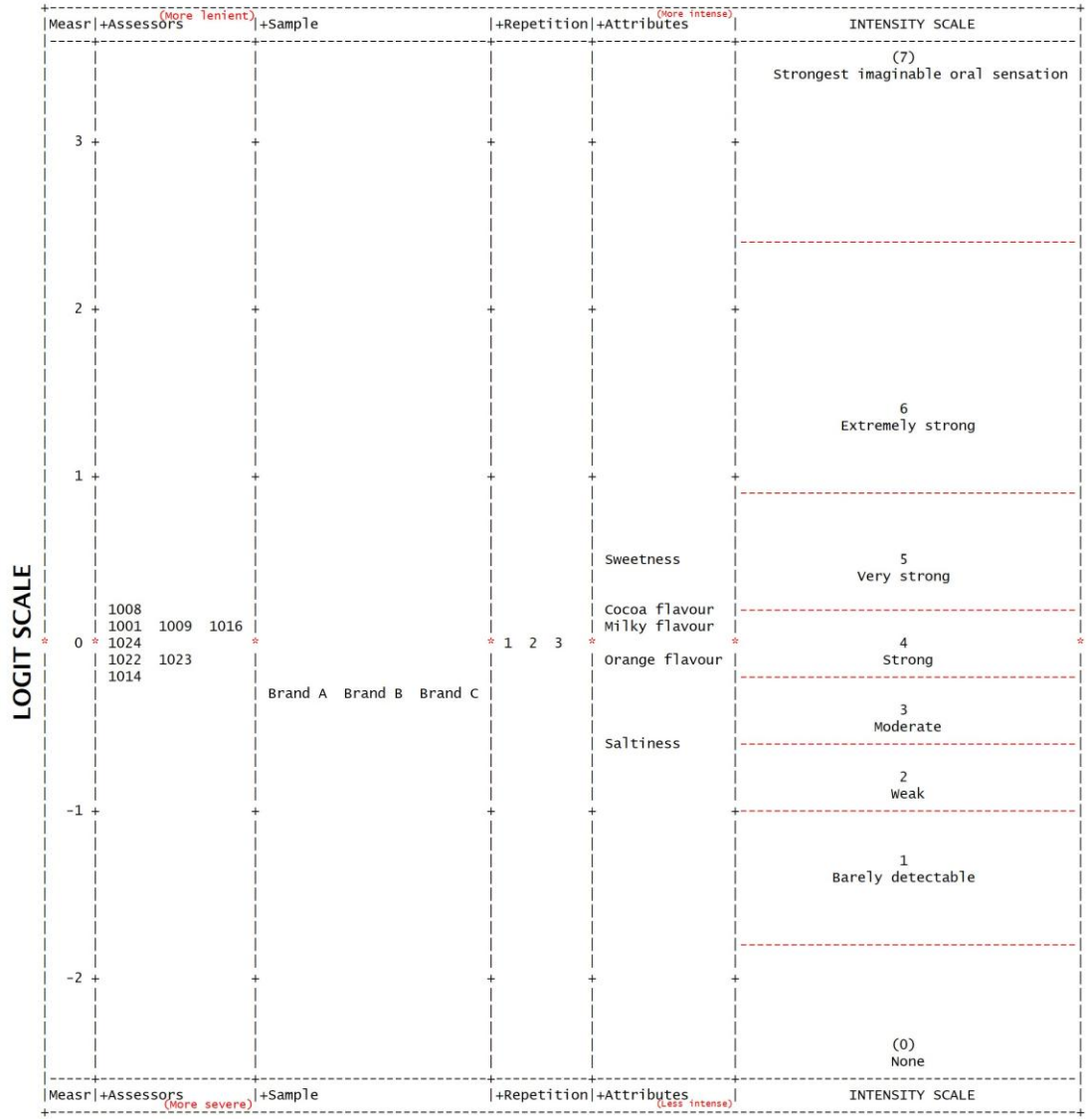


Figure 5.11. Many-Facet Wright Map for the Selected Untrained Assessors

The Wright map for the selected assessors (**Figure 5.11**) closely resembled that of the original untrained panel. However, the selected assessors displayed more consistent severity levels, with their measures tightly clustered around the mean (−0.2 to 0.2 logits) suggesting they applied the scale more uniformly after accounting for measurement error. The sample measures were also more tightly clustered, falling within a narrow range of 0.1 logits and remained below average based on the TIM, with no significant differences observed between the samples or their replicate evaluations on average. *Orange flavour* was still located below the average logit reflecting the continued impact of lower ratings for Brand B on the attribute. This was particularly evident as the selected assessors were able to

discriminate between the samples based on *Orange flavour* and were more accurate in rating the absence of *Orange flavour* for Brand B.

Notably, the half-point thresholds and the locations of assessors and attributes on the Wright map remained unchanged. This stability demonstrates the Rasch model's invariance property, where parameter estimates stay consistent across different subgroups, provided the model fit is adequate and the subgroups exhibit similar measurement characteristics (Bond et al., 2020). This consistency enables meaningful comparisons, even when panel composition varies, highlighting the robustness of the Rasch framework for sensory data analysis.

The Rasch model separation statistics and sample measures based on the Total Intensity Measure (TIM) are summarised in **Table 5.7**. The samples did not differ significantly, as indicated by a high chi-square (χ^2) p-value of 0.98, a strata value near 0, and a reliability of 0, confirming a lack of distinct levels between samples. The severity levels of the assessors were also not significantly different at a 95% confidence interval. Assessor severity levels also showed no significant differences at the 95% confidence level. Although the strata value of 1.45 suggested minimal variation among assessors, the reliability value of 0.41 indicated this variation was likely due to measurement error. Reliability values below 0.50 suggest that differences between measures are primarily due to measurement error (Wright & Masters, 2002).

Table 5.7. Summary of Rasch model separation statistics for the panel of selected untrained assessors

Rasch separation statistics	**Samples**	**Assessors**
Fixed χ^2 p-value *(α=0.05)*	0.98	0.06
Strata	0.33	1.45
Reliability	0.00	0.41
Sample Measure[1]		
Brand A	-0.32	
Brand B	-0.33	
Brand C	-0.32	

[1]Standard error (S.E) of 0.06 for all sample measures

5.3.4.3.1 Attribute contributions to the product differences

The importance of each attribute in differentiating the products is presented in **Figure 5.12** below. The acceptable range for Outfit Mnsq values is between 0.66 and 1.33, based on 72 responses per attribute. *Orange flavour* and *Milky flavour* emerged as key discriminating attributes, exhibiting underfitting Outfit Mnsq values of 1.56 and 1.15, respectively. Unlike the original panel, where *Cocoa flavour* had a higher Outfit Mnsq value (**Figure 5.5**), *Cocoa flavour* is now at the lower limit of the acceptable range (0.66), suggesting it did not effectively discriminate between samples. *Sweetness* now showed a higher value of 0.70, but both attributes were close to the overfit threshold, indicating limited contribution to product differentiation.

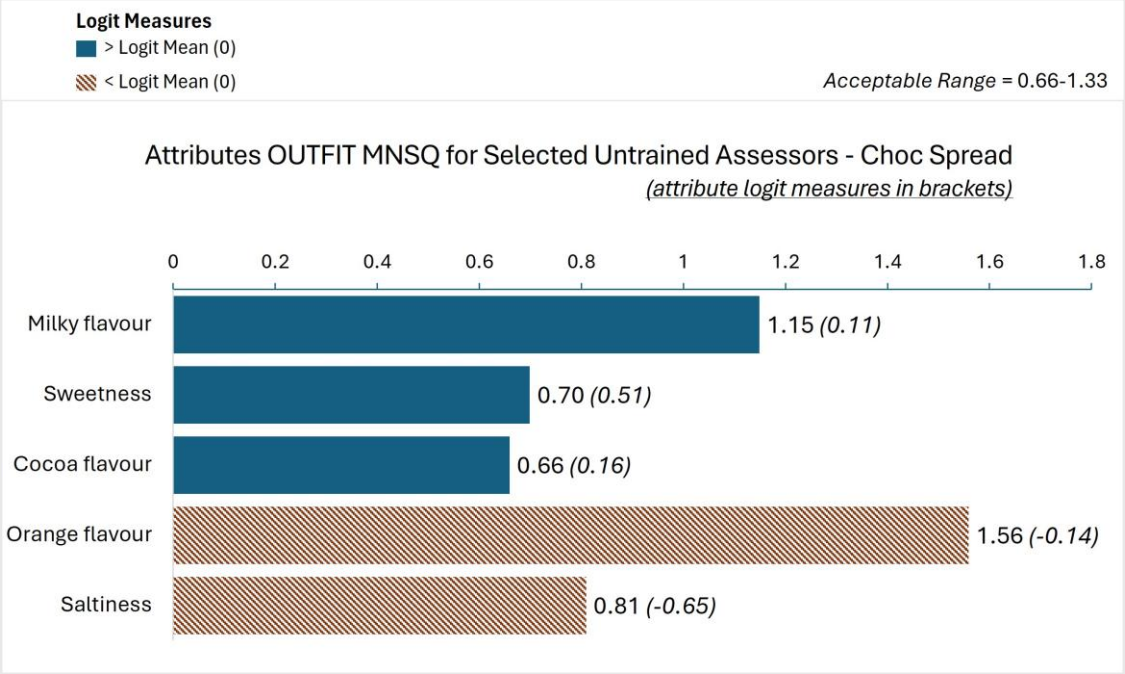


Figure 5.12. Attribute contributions to overall product differences for the panel of selected Untrained assessors

The slightly higher value for *Sweetness* may reflect differences in replicate interaction patterns within assessors. As shown later in **Figure 5.14** , only assessors 1016 and 1024 consistently ranked the samples based on *Sweetness*. However, all assessors, including these two, showed crossover interactions for *Cocoa flavour*. These rating patterns likely resulted in the loss of product discrimination for the attributes (Stone et al., 2012), and suggested that assessors struggled to consistently distinguish samples based on these attributes.

The intensity of *Orange flavour* remained lower than average (logit measure -0.14) because the extremely low ratings for Brand B, pulled down the overall average for *Orange flavour* across the samples. *Saltiness*, with a logit measure of -0.65, proved more challenging to rate, exhibiting an underfit (OUTFIT Mnsq = 0.81) for the selected assessors due to its low intensity and inconsistent ratings. These patterns are clearly reflected in the trellis plots for the selected assessors in **Figure 5.14**.

5.3.4.3.2 Relative performance of the selected assessors

In **Figure 5.13**, the control plot for the OUTFIT Mnsq values of the selected untrained assessors revealed no misfitting assessors, indicating that the rating patterns of all assessors were consistent with the panel's overall ability, (i.e., based on standard panel performance criteria, they showed panel agreement). However, their rating patterns (**Figure 5.14**) were generally more erratic, compared to the trained panel (**Figure 5.8**), though less so than those of the original untrained panel (**Figure 5.10**), particularly for the key attributes.

Assessors whose OUTFIT Mnsq values approached the acceptable limits tended to show stronger rating effects than others, as seen with Assessor 1001, who showed restriction of range effect across most attributes except *Orange flavour*. This supports the idea that the magnitude of a rating effect corresponds to changes in OUTFIT Mnsq values, which increase, or decrease based on how prominently an assessor displays that effect. As Linacre (1995) states, greater variance among parameters within a facet leads to higher OUTFIT Mnsq values. Similarly, lower variance corresponds to lower OUTFIT Mnsq values. This pattern is explored further in **Chapter 6**.

Single rater – rest of rater (SR/ROR) correlations were still inconsistent and therefore not as informative as with the trained panel as no clear pattern could be identified.

The selected untrained assessors were indeed the better performers within the untrained group, as identified using Rasch diagnostic tools, specifically the OUTFIT Mnsq control limits. The model identified the same key sensory attributes across both the trained and untrained panels.

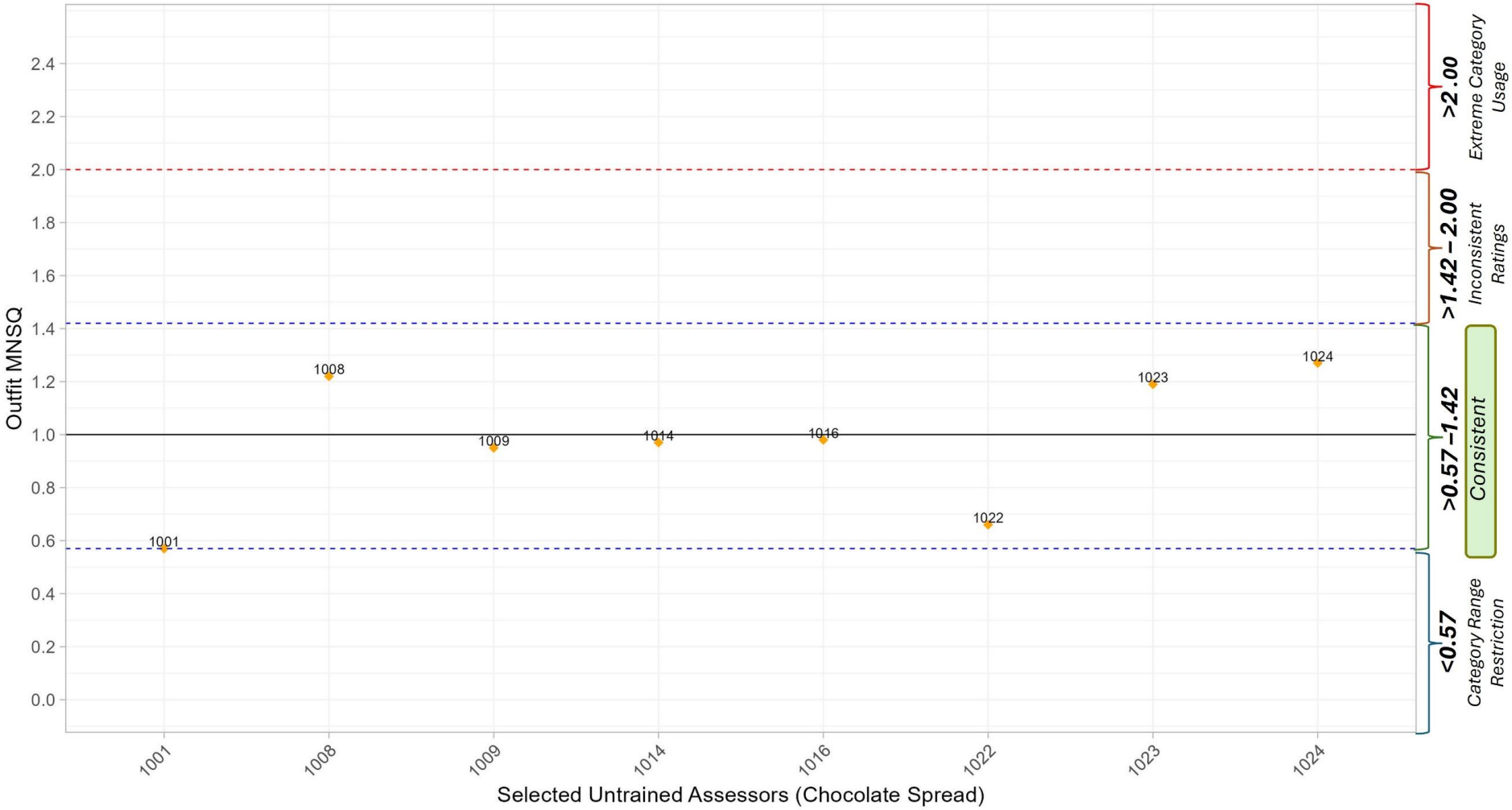


Figure 5.13. OUTFIT Mnsq plot for Selected Assessors from the Untrained panel

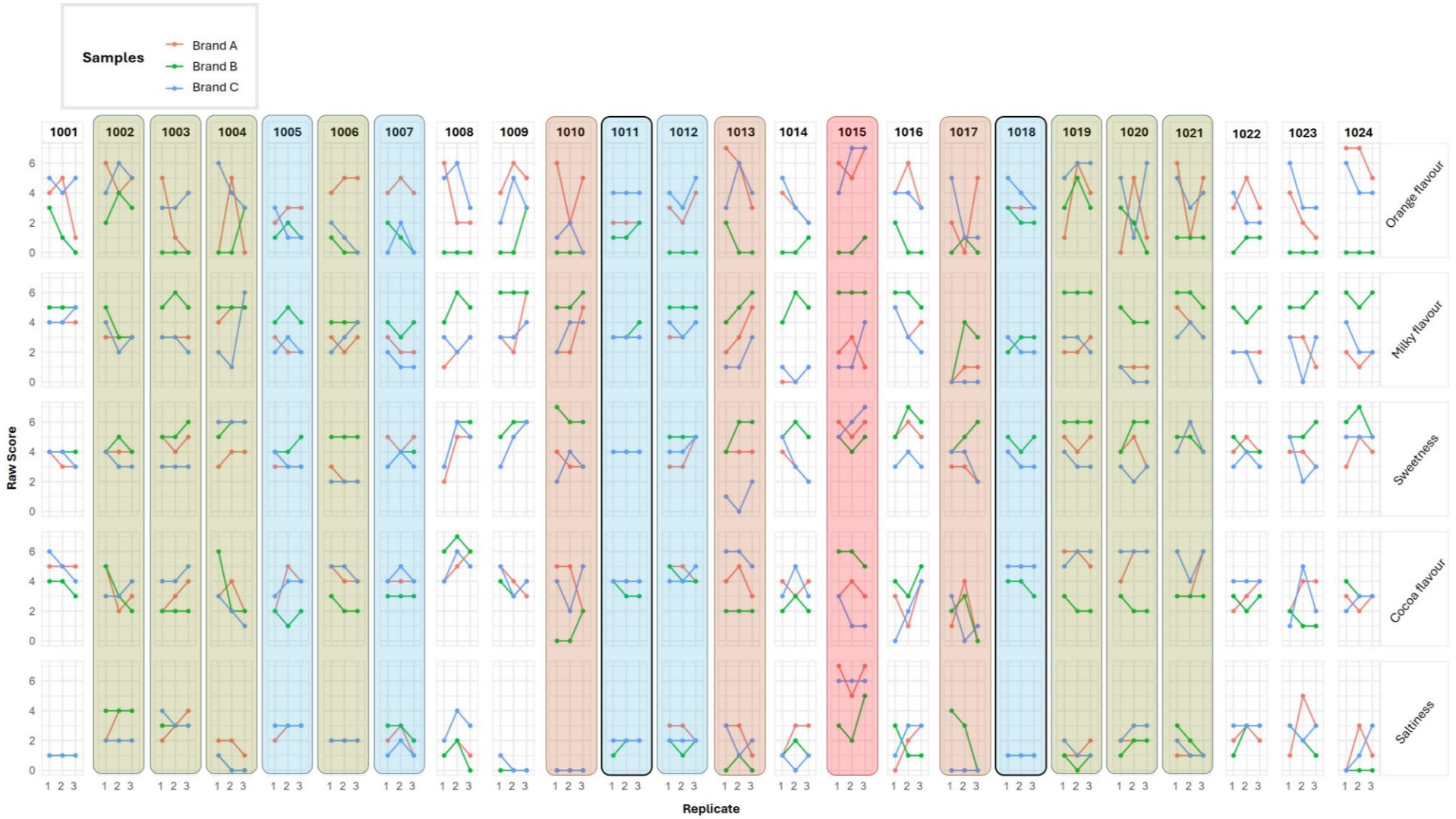


Figure 5.14. Trellis plots highlighting the Selected Assessors (**Unshaded**) from the Untrained panel, with shaded areas indicating excluded assessors. Shading colours represent Rasch-based interpretations of rater performance: **red** indicates extreme category use; **blue**, restriction of range; **blue with black borders**, central tendency; **brown**, inconsistent ratings; and **green**, assessors who were consistent based on Outfit Mnsq values but did not discriminate key attributes in the ANOVA.

However, while the trained panel was able to effectively discriminate between the products based on the Rasch measure of overall difference – TIM, neither the full untrained panel, nor the selected subset could differentiate the products holistically, though they could detect product differences in specific attributes. The untrained assessors rated the absence of *Orange flavour* in Brand B more accurately, whereas the trained panel, though rating it lower than the other attributes, appeared more conservative and did not rate it as completely absent. Still, the inconsistent and erratic ratings from the untrained panels, both for *Orange flavour* and other attributes, led to a loss of discrimination between products when considering the overall differences.

5.3.5 Convergence analysis of panel size on product discrimination

Convergence analysis was used to examine whether the untrained panel's lower discrimination reflected insufficient panel size or inconsistent rating patterns, and to compare how discrimination ability varied with panel size between trained and untrained assessors, following the method described in section **5.2.5.2: Convergence analysis**.

Figure 5.15 shows the *Sample facet* fixed chi-square values for both panels across two iterations. The untrained panel subsets (orange lines) produced low, irregular, and non-significant chi-square values ($p>0.05$) at all panel sizes, ranging from 0.1-0.4 at n=7 up to only 1.6 at n=24. Neither iteration reached significance, and the curves did not approach the discrimination level of the benchmarked original trained assessors. In contrast, the artificially expanded trained panel dataset (blue lines) produced a smooth, increasing discrimination trend and consistently high, significant chi-square values ($p<0.05$), increasing from 8.5-11.6 at n=7 to 24.6 at n=24. Both iterations followed almost identical patterns. This aligns with Myford and Wolfe's (2004) observation that when most raters in a Many-Facet Rasch analysis provide erratic ratings, the respondents being rated (in this case, the products) appear to differ only minimally in performance level, thereby reducing the ability to make reliable distinctions.

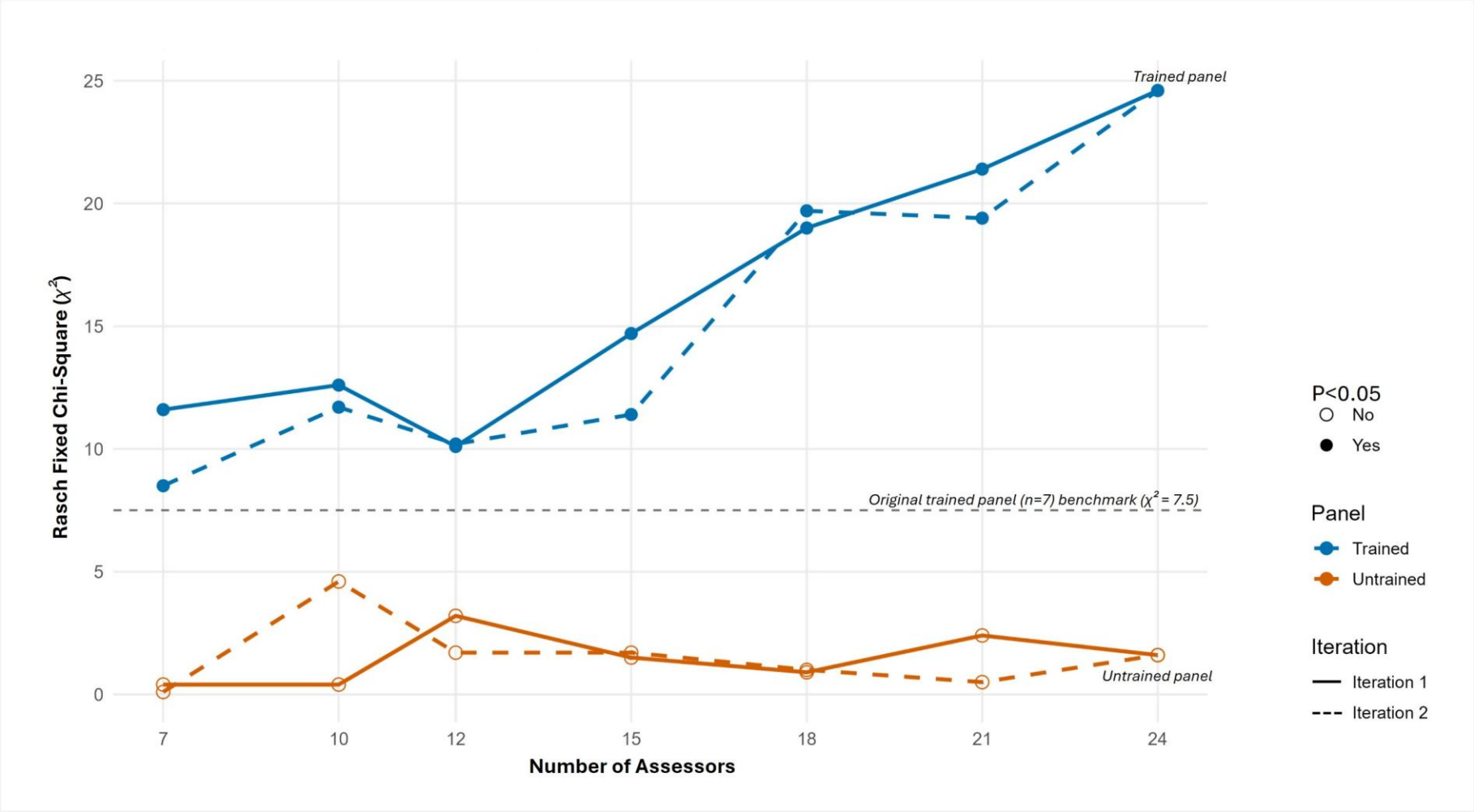


Figure 5.15. Rasch model fixed Chi-square convergence with increasing panel size for trained and untrained panels.
Trained-panel subsets (blue lines) show a consistent increase in *Sample facet* chi-square values as panel size increases, with discrimination stabilising beyond approximately 15 assessors. Untrained-panel subsets (orange lines) display irregular and generally low chi-square values, indicating poor discrimination and lack of convergence relative to the trained panel.

The trained panel's convergence curve suggested that discrimination power increases in a stable, predictable way when underlying ratings are consistent. The untrained panel's failure to converge indicates that rater inconsistency, rather than insufficient sample size, limited the model's ability to detect product differences. Minor fluctuations in the trained panel curve at small subset sizes likely reflect the influence of individual assessors on model calibration, with stability reached at around 15 assessors.

Another possible contributor to the untrained panel's lack of discrimination is the opposing attribute directions for Brand B, which was the only sample without added orange flavouring. The untrained assessors rated Brand B lowest for Orange flavour but highest for Milky flavour. When the Rasch model adjusted these ratings to estimate a single latent Overall difference variable, the opposing directions may have partially cancelled each other out, reducing apparent product separation on the logit scale. Although the trained panel showed similar rating patterns for Brand B, their lower measurement noise on the less dominant attributes allowed for clearer discrimination between the samples.

This highlights an important caveat in applying the MFRM to estimate an overall difference construct. The present results suggest that when a product exhibits opposing attribute intensities (e.g., low on one attribute but high on another), variability in ratings can cause the combined latent estimate to mask genuine sensory differences. This was evident for Brand B, which lacked added orange flavouring but was the only sample containing milk chocolate crumbs and full cream milk (**Table B 2**), a formulation likely reflected in its perceived attribute intensities. However, because no analytical tests were conducted to verify the flavour composition of the samples, this interpretation should be viewed as tentative.

These findings suggest that the model is most appropriate when products being examined do not have extreme opposing attribute profiles and when assessors provide relatively consistent ratings, even if severity levels differ. The advantage of the MFRM over traditional methods is its integrated diagnostic framework: it simultaneously evaluates product discrimination, adjusts for systematic assessor severity and leniency, and identifies problematic raters within the same analysis.

However, this adjustment only works for consistent bias (e.g., an assessor who is always more severe). It cannot correct erratic rating patterns or differences in attribute conceptualisation. Analysts also need to consider test design carefully, as unidimensionality is ultimately driven by which attributes are selected to represent the underlying construct.

When products are expected to exhibit opposing attribute profiles, examining attribute-level differences individually through the Rasch bias/interaction plots can reveal contrasts that become obscured when scores are collapsed into a single measure. As shown in **Figure E 2,** these plots illustrate the opposing attribute intensities and highlight product differences evident in the rating patterns that are not visible in the aggregated scores.

Finally, while the convergence analysis shows that increasing panel size could not compensate for a lack of training in this study, MFRM may still offer value for trained panels by accounting for residual severity differences and potentially reducing the amount of recalibration needed to maintain panel alignment.

5.4 Limitations of the study

Panel performance constraints

Although untrained assessors are expected to exhibit inconsistencies due to factors such as lack of expertise, adaptation, or poor sensitivity, the poor scoring repeatability observed in many of them may have been influenced by carryover effects. Unlike the trained panel, the untrained assessors completed all evaluations in a single session without specified time gaps between replicate evaluations, which likely contributed to fatigue or reduced focus.

Motivation may have been an additional contributing factor. The absence of incentives could have lowered engagement, whereas reward systems and feedback are known to enhance assessor enthusiasm by reinforcing the perceived value of their contribution (Findlay et al., 2007; Kemp et al., 2018; Meilgaard et al., 2025). Collectively, these factors may have hindered the untrained assessors' ability to generate reliable and consistent data, particularly in contrast to the trained panel, who were accustomed to the demands and importance of sensory evaluation tasks.

Lack of scale-use training

A further limitation affecting both panels was the absence of training on scale-use which could have contributed to the product interactions observed within the trained panel and the more erratic ratings seen among the untrained assessors, alongside possible differences in sensory sensitivity and attribute interpretation. Although the Rasch model can statistically accommodate individual differences in severity, where some assessors consistently rate higher or lower than others (Linacre, 1994; Myford & Wolfe, 2004; Bond et al., 2020), training on scale interpretation and calibration could have reduced these variations and improved measurement precision. This represents a readily addressable limitation for future studies, where targeted training on how to apply the scale could enhance assessor consistency and reduce measurement error without requiring all assessors to adopt identical rating patterns.

Lack of instrumental analysis to verify product characteristics

Another limitation was the absence of instrumental analysis to objectively quantify sample composition. Sensory data indicated that Brand B was rated highest in *Milky flavour* and lowest in *Orange flavour*, consistent with its formulation (the only product containing milk chocolate crumbs and full cream milk, and the only one without added orange flavouring). However, without analytical tests to verify actual concentrations of dairy or citrus-related compounds across the three sampled brands, it is not possible to confirm whether these perceived attribute intensities reflect true chemical differences or perceptual interactions among ingredients. Incorporating instrumental analysis (e.g., GC-MS for volatile compounds or HPLC for non-volatile flavour components) in future work would strengthen interpretation of sensory attribute patterns.

Sample choice constraints

Finally, while choosing a sample with no orange flavouring to compare against other chocolate-orange spreads was intentional, the opposing attribute intensities it produced were not anticipated to cause cancellation effects in the MFRM overall difference estimation. This limited the possibility of making meaningful comparisons with the Jaffa cakes study (Chapter 4), in which products were

deliberately selected to share similar sensory attributes and thereby support cross-study evaluation of MFRM performance across different food matrices.

While these limitations influenced panel performance and constrained cross-study comparisons, they also clarify where methodological refinements can enhance the robustness of future sensory research using the Many-Facet Rasch Model (MFRM).

5.5 Significance of the study

In this study, the conventional ANOVA approach to evaluating assessor and panel performance provided useful insights into the discriminatory abilities of the assessors and the panels. Typically, the discriminatory ability of assessors is evaluated based on whether they detect differences in the key attributes identified by the panel, with those failing to do so selected for retraining. However, this study revealed some limitations in that approach. The trained panel ANOVA results showed that four out of the five sensory attributes (except *Saltiness*) exhibited significant differences between products ($p<0.01$), identifying them as key attributes. While *Orange flavour* and *Sweetness* showed significant interaction effects, which may affect their reliability. As a result, *Milky flavour* and *Cocoa flavour* were identified as more stable key attributes. However, individual ANOVAs revealed that no assessors significantly discriminated between products on *Milky flavour*, and only two out of seven did so for *Cocoa flavour* ($\alpha = 0.05$), indicating a misalignment between panel-level and individual-level findings. Closer examination of the raw data using the trellis and interaction plots confirmed this inconsistency, emphasising the importance of closely interrogating raw data when evaluating panel performance. As previous research has noted, the ANOVA alone does not provide all relevant diagnostic information (Tomic et al., 2007; Stone et al., 2012; Ho, 2015; Raithatha & Rogers, 2018). This disconnect between the panel and individual results can limit its utility for selecting key attributes, especially in situations where discriminatory attributes are not known a priori, thereby reducing the efficiency of the method for guiding assessor selection and training.

In contrast, Rasch analysis provided more diagnostic and interpretable insights into panel performance without requiring multiple statistical tests. By first adjusting for assessor severity, it removed a major source of individual variability in sensory data

(i.e., scale level effects), so the remaining variance more accurately reflected true product differences provided that assessors rated the attributes consistently, as was the case for the trained panel. Interaction effects from other modelled variables (i.e. *Repetition* and *Sample* facets) and crossover variations were captured as measurement errors, which were reflected in the residual diagnostics indices (OUTFIT Mnsq). Consequently, the OUTFIT Mnsq rankings for each attribute were primarily influenced by variations in rating patterns across samples. This approach allowed for a more precise identification of key attributes and proves potentially beneficial in situations where there is no prior knowledge of expected differences. This clarity was further enhanced when Rasch outputs were combined with response distribution and interaction plots, which revealed additional insights, such as where loss of discrimination by an attribute was a result of crossover interactions.

Additionally, correlation analysis of the standardised residuals allowed for the identification of locally dependent attributes, those whose ratings were statistically influenced by other attributes. This provided valuable diagnostic insight into subtle product nuances that might otherwise go unnoticed. In practice, this information can guide the combination or redefinition of attributes for clearer, more reliable sensory profiling. In this study, the addition of milk chocolate not only influenced perceptions of milky and cocoa flavours but also altered the perceived *Sweetness*. *Sweetness* and *Milky flavour* were flagged as locally dependent attributes across the trained panel, the full untrained panel, and the selected subset of untrained assessors, although the degree of dependence varied across the different panels. Recognising and accounting for local dependence ensures that the measures reflect meaningful sensory differences, rather than overlapping perceptions that could overstate the distinctiveness of individual attributes.

At the individual performance level, after adjusting for differences in assessor severity (as discussed in section **3.1.2**), Rasch diagnostics (OUTFIT Mnsq and SR/ROR correlation), along with distribution plots of the raw data, revealed assessors who rated the samples in an order that differed from the rest. These diagnostics also highlighted rating effects such as restriction of range, central tendency, inconsistent scoring, and extreme category usage. While the Rasch

model clearly offers advantages for identifying assessor bias, rating effects, and local dependence, aspects that ANOVA does not easily detect, it does not replace the ANOVA.

ANOVA provides a familiar framework for statistical significance testing but assumes homogeneity of scale use, and does not account for individual rating behaviours, which can obscure true product differences or lead to inefficient assessor selection if discriminatory attributes are not known in advance. In contrast, the Rasch model adjusts for these individual differences and provides additional diagnostic tools that improve data quality and interpretation. By combining Rasch diagnostics with insights from individual ANOVA results, which identified assessors capable of discriminating between samples, better-performing assessors could be selected from an untrained panel.

Just like an *X-ray* that detects problems early and precisely, the Rasch model enables panel leaders to pinpoint where there are issues in assessor performance, allowing potential concerns to be flagged before more statistical analysis is required. While mean-based interaction plots can also reveal inconsistencies and crossover patterns among assessors, the Rasch model provides a quantitative and model-based assessment of fit, estimating how well each assessor's ratings align with the expected response pattern. This complements the descriptive information provided by the plots. Tools such as Wright maps and OUTFIT Mnsq plots provide a rapid, visual overview of individual assessor behaviour, and how consistently attributes are rated across the panel. This could be useful for identifying individuals from untrained panels with potential for recruitment into expert panels, or for detecting subtle declines in performance or panel drift within trained panels. Additionally, it highlights attributes that may require further training to ensure assessors rate them consistently and accurately. It also supports decisions on whether certain attributes are worth further measurement, or if they should be reconsidered as redundant and removed from the panel evaluation.

It should be noted, however, that Rasch fit values are relative to the response pattern within the panel, not to an ideal standard. Therefore, acceptable fit in a poor-quality panel does not mean good performance. In panels where overall performance is inconsistent like in untrained panels, acceptable fit values may be

misleading because the model is benchmarking them against a weak reference group. Therefore, fit statistics should be interpreted alongside raw data visualisations and other performance indicators to ensure accurate assessment of assessor performance. In line with earlier observations on the importance of always examining the raw data, Myford and Wolfe (2004) similarly advise that researchers inspect the vectors of observed ratings for any overfitting or misfitting assessors before concluding that apparent rater effects, such as central tendency or severity, represent genuine behavioural differences rather than artefacts of the data.

However, the true strength of the Rasch-based approach lies in its ability to complement, rather than replace, traditional sensory quality control methods. When integrated with ANOVAs, raw data distribution plots and interaction plots, the Rasch model enhances the analytical capabilities of sensory analysts, reducing the need for extensive statistical testing and resource use. This combination offers a more comprehensive, multi-layered understanding of panel performance, merging intuitive visual insights with robust statistical analysis and individual diagnostics.

Ultimately, while the Rasch model adjusts for systematic differences in assessor severity, allowing for individual rating tendencies rather than requiring unanimous panel ratings, (Linacre, 1994), it cannot compensate for fundamental issues such as poor attribute understanding, low sensory sensitivity, or erratic scale use. As noted in the limitations, none of the panels received formal training on scale use, which likely contributed to the variation observed within the trained panel and the inconsistency seen among the untrained assessors, in addition to possible differences in sensory sensitivity and attribute interpretation. Despite the limited overall reliability of the panels, the Rasch diagnostics still identified differences in assessor severity and fit, consistent with the expectation that individual differences in scale use (e.g., severity, range restriction) introduce additional variability. Therefore, these results support the conclusion that although Rasch analysis can adjust for level effects statistically, adequate training remains essential to ensure consistent attribute interpretation and rating precision.

The Rasch model was primarily used as a diagnostic tool to identify individual rating behaviours and highlight areas for improvement; however, future work could use these individual standards to track each assessor's consistency over time, ensuring

that stable individual patterns are retained while problematic drift or bias is identified early. Complementary to conventional methods, Chapters 4 and 5 have demonstrated how the Many-Facet Rasch Model (MFRM) provides an efficient diagnostic approach for examining panel performance, delivering deeper insights into product differences and the attributes perceived to drive those differences. The Rasch-based approach supports high standards in sensory quality control by statistically adjusting for individual scale-level effects (i.e., consistent severity or leniency differences among assessors), which could reduce the reliance on intensive training to force all assessors to use the scale identically. By modelling each assessor's severity and consistency, the Rasch model transforms raw scores to a common interval scale, ensuring that valid product comparisons can still be made even when assessors apply their own consistent standards. This offers the potential for significant cost savings in both time and resources, improving overall efficiency in sensory quality control without compromising the quality of insights gained.

Building on the preceding discussion, the next chapter examines the potential of the MFRM to strengthen sensory quality diagnosis by integrating product evaluation and assessor performance within a unified analytical framework. While this chapter highlighted that Rasch fit values (i.e., OUTFIT Mnsq) are relative to the response patterns within a panel and must therefore be interpreted cautiously, the next study explores how these diagnostics can still support practical decision-making in sensory quality programs when used alongside raw data and other performance indicators.

Specifically, the model is applied to sensory data from an untrained panel assessing a different product with a broader set of sensory attributes. The analysis investigates whether identifying assessors who show more stable response patterns relative to their group (i.e., acceptable fit within that context) can improve panel discrimination and reliability.

This exploratory work does not treat fit as an absolute measure of assessor competence but instead examines whether Rasch-based diagnostics, applied critically, can inform assessor selection and recruitment in sensory quality control settings.

Chapter 6

A Unified Rasch Approach to Sensory Difference Testing and Quality Control: A Validation Study

6.1 Overview

To evaluate the robustness and transferability of the Rasch-based approach for sensory difference testing, both the assessment of overall product differences and the monitoring of assessor performance were applied in a new context, with the aim of validating the framework as a unified tool. As demonstrated in previous chapters, the Many-Facet Rasch Model (MFRM) allows for simultaneous evaluation of both product differences and assessor performance within the same analysis, providing deeper insights on a granular level.

In this chapter, the MFRM is used to assess product differences and assessor performance, while Principal Component Analysis of Residuals (PCAR) is employed to uncover patterns and relationships between sensory attributes that may be overlooked by conventional methods. The results are interpreted with practical implications for sensory quality control, including confirming product differences without confounding from individual differences in scale use, evaluating assessor reliability, identifying outlying assessor behaviour, selecting better-performing assessors, guiding training needs, determining the contribution of attributes to product differences, and assessing the utility versus redundancy of attributes.

The chapter also addresses key limitations from Chapter 4, where the choice of attributes used in the Rasch-based TIM approach was not fully representative of the perceptible attributes in the products as perceived in a DFC test. To resolve this, a preliminary sensory session was conducted to select sensory attributes, as is standard practice with attribute difference testing in sensory quality control. Although untrained panels were still used, the participant information document and preliminary instructions (Appendix **C.3**) were designed to emphasise the need for a high level of commitment, aiming to address the limitation from Chapter 5, where the untrained panel seemed less motivated compared to the trained panel.

This study used DFC and attribute intensity ratings on tomato soup samples.

6.1.1 Objectives

The aim of the study was to evaluate the robustness and transferability of the Rasch-based framework for sensory difference testing and quality control, applying it to both the measurement of overall product differences and monitoring assessor performance, all within the same analysis, but in a new context, while also addressing limitations identified in the previous studies.

The specific objectives were:

1. To identify perceivable attributes in the tomato soup samples for use in the attribute intensity rating (AR) test of the main study.
2. To evaluate the overall difference between three tomato soup samples using the DFC test.
3. To assess the intensities of identified sensory attributes (from objective 1) in the tomato soup samples using the AR test.
4. To estimate the Total Intensity Measures (TIM) by combining the intensity ratings from the identified attributes using the MFRM.
5. To compare the overall difference results from the DFC ratings with the TIM derived from the combined attributes, using pairwise comparison tests.
6. To assess the performance of the untrained panel in rating the tomato soup samples and select the top-performing assessors based on Rasch model residual fit statistics (OUTFIT Mnsq).
7. To investigate assessor rating behaviour in relation to Rasch model residual fit statistics (OUTFIT Mnsq).
8. To identify the key attributes responsible for the differences between the tomato soup samples.

6.1.2 Study highlights

- Significant differences were observed between the tomato soup samples and the control sample.

- The MFRM Wright map provided a clear visual summary of the dataset, showing assessor severity levels and identifying the attributes most strongly perceived across the samples.
- PCAR revealed response dependencies between attributes, including both expected correlations (e.g., *Rich aroma* and *Savoury flavour*) and less conventional ones (e.g., *Herby appearance* and *Viscous appearance*).
- Rasch group-level statistics revealed inconsistencies in the application of the rating scale, both by the full panel and the selected assessors.
- OUTFIT Mnsq ranges for assessors showed specific rating effects, which were aligned with the pattern of the raw rating scores observed in the trellis plots.
- Key and redundant attributes were identified through the OUTFIT Mnsq, with *Creamy flavour*, *Thick mouthfeel* and *Viscous appearance* driving the most significant differences, while *Cooked tomato* characteristics and *Colour intensity* were found to be the most redundant.
- *Creamy flavour* and *Rich aroma* were among the most challenging attributes for the assessors to evaluate, as indicated by the OUTFIT Mnsq.

6.2 Sensory study: materials and methods

Sensory data were from the dataset referenced here as (Ariakpomu et al., 2025a).

6.2.1 Samples

Tomato soup was selected for this study due to its versatility and widespread familiarity, making it appealing for assessors to evaluate. Its ease of modification also allowed for the creation of samples with varied sensory characteristics and attribute intensities, enabling clear hypotheses about expected differences. Two types of ready-made canned tomato soup, “cream of tomato soup” and “cream of tomato and basil soup”, were used as the base, and the three final samples were prepared by modifying these bases with additional ingredients. This modification was also important to mask the original flavours of the base products, reducing the likelihood of bias from assessors who might have easily recognised them, as previously observed with the Jaffa cakes study in **Chapter 4 (pg. 86)**. The cream of tomato and basil soup was used in its original form as the reference sample.

To create the second sample, double cream and dried chopped basil leaves were added to a cream of tomato soup base, aiming to increase its creamy flavour and thickness while aligning it more closely with the basil flavour of the reference sample. For the third sample, the cream of tomato and basil soup base was modified with the addition of passata and garlic granules to enhance its aromatic and savoury characteristics relative to the reference. Consequently, the hypothesis was that Sample 2 would be perceived as the thickest and most intense in creamy flavour while having a similar herby flavour to the reference sample, and Sample 3 would have stronger savoury and aromatic notes compared to sample 2 and the reference sample. These expected sensory differences formed the basis for the product comparisons.

All soup bases, along with the additional ingredients (passata, garlic granules, and dried chopped basil leaves), were purchased from a UK retail store and stored at room temperature (20±3°C). The double cream, also purchased from the same store, was stored separately in the refrigerator at 4°C until sample preparation. The ingredients were then incorporated into their respective soup bases in specified proportions, as outlined in **Table B 3**. The soups were heated in saucepans over medium heat on a stove top, with occasional stirring, until they began to gently bubble. After heating, they were allowed to cool to a serving temperature of approximately 70°C before being transferred to insulated flasks, accounting for potential heat loss during serving. The final samples were served to assessors within a temperature range of 60-67°C throughout the course of the testing sessions each day, ensuring realistic consumption conditions and minimising unexpected bias.

6.2.2 Participants

Ethical approval for the sensory study was granted by the Business, Environment and Social Sciences Faculty Research Ethics Committee at the University of Leeds. Participants (n=54) all residents of Leeds, and the majority being staff or students at the University, were recruited through emails, poster advertisements and personal referrals. Participants were eligible if they were aged between 18 and 65 years, did not have any chronic health conditions, were not allergic or intolerant to the ingredients in the tomato soup samples or the palate cleanser, were not taking any routine medication (with the exception of contraceptives), were not following

any special or restricted diets, were not pregnant or lactating, and were available to attend two 1-hour sensory testing sessions within one month, with a minimum interval of four days between sessions.

Each participant was provided with detailed information about the study requirements, including data protection and the data sharing disclaimer. Informed consent was obtained through signed consent forms, completed both at the point of enrolment and again in hard copy upon attendance at the first study session, to ensure participants fully understood the study requirements and were willing to proceed. Detailed instructions for the sensory test procedure were sent within three days of the scheduled session, with reminders sent at 1 hour and at 15 minutes before the session. These reminders were programmed as *in-person* Microsoft Teams meetings, which were automatically added to participants' calendars. The decision to use Teams was made to improve attendance and punctuality, as the previous study had experienced issues with scheduled participants failing to attend, arriving late, or missing the second part of the test. As a result, all 54 participants in this study attended both test sessions and were generally punctual.

The final untrained panel consisted of 35 females (65%) and 19 males (35%), aged between 18 and 54 years. They represented various ethnicities: 16 Asian (30%), 10 Black (18%), 17 White (31%), 3 Mixed (6%), and 8 belonging to other ethnic groups (15%). All assessors, except one, reported consuming soup products at least a few times a year. 40 participants (74%) consumed tomato soup at least a few times annually, and 30 (55%) had previous experience participating in sensory evaluation tests.

As in the previous study, participants selected two convenient test dates via an online form (Jotform Inc, 2023). Upon completion of both sessions, each assessor received a £20 Flexi Gift voucher (GiftPay, 2024) as an incentive for their participation.

6.2.3 Study design

As in **Chapter 4**, a Randomised Complete Block Design (RCBD) and Latin Square were used to account for order effects and other potential sources of variation in the sensory experiments. Each assessor participated in two separate sessions: one for the DFC test and another for the AR test, with a minimum gap of four days

between the sessions. To minimise expectation biases (Meilgaard et al., 2015), half of the participants completed the AR test first, while the other half started with the DFC test. Additionally, to reduce variations due to the time of day, participants were only able to select either two morning sessions or two afternoon sessions for both tests, with the appointment booking form automatically ensuring this balance. Attendance was carefully managed to ensure a balance between the time of day and the order in which participants completed the tests.

In each test session, three samples were presented. For the AR test, the samples were presented monadically (one at a time), while for the DFC test, the samples were presented in pairs, consisting of a test sample and a reference sample. Each sample was evaluated three times, resulting in a total of nine evaluations for the AR test and eighteen evaluations for the DFC test. All samples were served warm, with temperatures ranging from 60 to 67°C, in 30ml clear plastic shot cups labelled with random three-digit codes. The reference sample for the DFC test was labelled "R".

A limitation of the study discussed in **Chapter 4** was that the attributes used for the AR test, which are combined to estimate the latent variable (TIM), did not fully represent the perceivable differences in the product, as no prior testing was conducted to select the relevant attributes. This oversight may have affected the results of the comparison between the DFC and AR tests. The DFC test assesses only overall product differences, meaning that attributes not included in the AR test could have still been perceived in the DFC test, potentially influencing the conclusions drawn from the comparison.

In this study, this limitation was addressed by conducting a preliminary evaluation session to identify the perceivable attributes across the samples of interest.

6.2.4 Attributes selection

Following the methods described by Lee et al. (2021), Giacalone and Hedelund (2016), and Zeppa et al. (2012) with slight modifications, sensory descriptors were generated by untrained assessors (n=7), three of whom had experience with descriptive analysis. The assessors were presented with the three tomato soup samples (described in section **6.2.1**), one at a time. For each sample, they were

asked to describe the sensory characteristics perceived, and to rate how well each sample exhibited those characteristics on a scale from 0 (not at all) to 5 (very well).

A total of 61 descriptors, both comprehensive and specific, were generated by the 7 assessors, with some descriptors occurring more frequently across the group. A final list of 18 attributes was generated, across 5 sensory modalities based on how often they were mentioned. Similar descriptors were consolidated into common terms. The selected attributes, along with their corresponding definitions, are outlined below.

Table 6.1 List of 18 sensory attributes across 5 modalities used for the AR test, including definitions

	Attributes	Definitions
Appearance	Glossy appearance	Degree of shine or reflected light from the surface (Tomaschunas et al., 2013).
	Herby appearance	The presence of small, chopped pieces of herbs.
	Colour intensity	Intensity or strength of colour from light to dark (Meilgaard et al., 2025).
	Viscous appearance	Thick and slow-moving when you tilt the container.
Aroma	Pungent aroma	Sharp, physically penetrating sensation in the nasal cavity.
	Rich aroma	Combination of multiple ingredients creating a deep and full aroma. E.g. well-seasoned food.
	Cooked tomato aroma	Typical smell of cooked tomato.
Mouthfeel	Smooth mouthfeel	Feels velvety or silky in the mouth, not rough or grainy (Cliff et al., 2013).
	Homogeneous mouthfeel	Feels the same way throughout.
	Thick mouthfeel	Feels dense or heavy in the mouth.
Flavour	Creamy flavour	Flavour associated with dairy products. E.g. cream, cheese.
	Savoury flavour	Rich, spicy flavour associated with vegetable or meat broth.
	Herbal flavour	Underlying flavour of dried herbs. E.g. basil, oregano.
	Cooked tomato flavour	Typical cooked tomato flavour.
Taste	Sweet taste	Typical sweet taste. E.g. sugar/sucrose.
	Sour taste	Sharp, tangy or tart taste. E.g. citric acid in lemons.
	Salty taste	Typical salt flavour. E.g. common salt / NaCl or seawater.
	Aftertaste	Residual taste in mouth after ingestion (Mitchell et al., 2011).

6.2.5 Sensory evaluation procedures

The sensory evaluation procedures used in this study were similar to those described in **Chapter 4** (as described in the following paragraphs), with adjustments made to suit the specific product type and to address previously identified limitations. To enhance time efficiency and ensure assessors were familiar with the process, a preview of the test instructions was sent to them ahead of their scheduled sessions. This allowed them to review the procedures ahead of time and arrive with a clear understanding of what to expect. In the previous study, where instructions were provided only on the day of testing, some assessors skimmed or skipped them entirely, which impacted the consistency of the evaluations. Copies of the preview instructions and the questionnaires used in the study are included in Appendix **C.3** and **C.4,** respectively.

For the DFC test, each assessor received 10ml of each sample and was informed that some coded test samples might be identical to the reference. They were instructed to drink directly from the sample cups and consume the entire contents at once, while assessing the overall sensory experience. This instruction was necessary to ensure that all assessors evaluated each sample fully, ensuring consistency in the sensory experience when comparing the test and reference samples. They were directed to first taste the sample labelled "R", then taste the coded test sample, and rate the size of difference perceived between them, using a unidirectional labelled 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.

After completing the third replicate evaluation, assessors were asked to reflect on the products they had evaluated and, on the following page, identify the reasons for any differences perceived. This section included yes/no questions listing all the attributes used in the AR test, along with their corresponding definitions, and asked assessors to indicate which attributes they perceived to be different between any of the samples, and the control sample (R). An additional comment section was included for assessors to note any other perceived differences not captured by the listed attributes. This addition was necessary to confirm whether the AR test captured all the relevant sensory attributes identified during the DFC test.

For the AR test, assessors rated the perceived intensities of the eighteen attributes listed in **Table 6.1** above. Specific instructions and definitions for each attribute were provided, outlining the exact procedure in which the samples should be evaluated across the five modalities. These instructions were summarised on the first page of the questionnaire for the first replicate evaluation (see **Figure C 11**). Specifically, assessors were instructed to first evaluate the appearance attributes, followed by the aroma attributes. For the oral evaluation, they were asked to take three sips of the sample: the first sip was for examining mouthfeel, the second for assessing flavour attributes, and the final sip for taste attributes. Each assessor was presented with 25ml of each sample and asked to rate the intensity of the attributes based on the instructions provided (outlined below with the corresponding attributes). Attribute definitions were included in brackets for each attribute within the questionnaire.

1. **Appearance:** Pick up the sample, examine it by looking directly into the cup and rate how strong the following appearance attributes are: *glossy*, *herby*, *colour intensity* and *viscous appearance*.

2. **Aroma:** Smell the sample and rate how strong the following aroma attributes are: *pungent*, *cooked tomato*, and *rich aroma*.

   These appearance and aroma attributes were listed on the same page of the questionnaire. On the following page, the mouthfeel, flavour, and taste attributes were presented, with the evaluation instructions as follows:

3. **Mouthfeel:** Take the first sip of the sample, and before swallowing, pay attention to how it feels in your mouth. While you assess the mouthfeel, rate how strong the following attributes are: *smooth*, *homogenous*, and *thick mouthfeel*.

4. **Flavour:** Take another sip, and before swallowing, focus on the different flavours noticed. While assessing the flavours, rate how strong the following attributes are: *herba*l, *creamy*, *savoury*, and *cooked tomato flavour*. The order of the flavour attributes was randomised for each sample and assessor, as suggested by (Ares et al., 2014) attempting to reduce errors of habituation, logical error and halo effect (Lawless & Heymann, 2010; Meilgaard et al., 2025).

5. **Taste:** take the last sip and move it around your tongue to fully experience the taste. While assessing the sample, rate how strong the following attributes are: *sweet*, *sour*, and *salty taste*, as well as *aftertaste*.

The same 8-point categorical intensity scale from the previous studies was used, ranging from 0 to 7 with labels adapted from the Labelled Magnitude Scale (LMS) (Green et al., 1996). 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.

Assessors were provided with a cup of water and plain crackers to cleanse their palate between sample evaluations. A mandatory interval was observed between each sample; 15 seconds for the DFC test and 30 seconds for the AR test, to ensure adequate palate cleansing. To minimise sensory fatigue and memory bias, assessors were also given 5-minute breaks between replicates for both tests.

6.2.6 Data analysis

Rasch and statistical analyses were conducted following the procedures described in section **3.3**. As in the previous studies, the Attribute Rating (AR) data were fitted to a Many-Facet Rasch Model (MFRM) with four facets: *Assessors*, *Samples*, *Repetition*, and *Attributes*, as detailed below (TIM). To enable comparison between the two approaches, a separate model was employed for the DFC data (DFCM). Unlike the previous study, where three separate datasets were created, one for each replicate evaluation, and separate models were fitted, the TIM and DFCM analyses in this study did not involve splitting the data. Instead, all replicate evaluations were retained within the datasets, and Repetition was explicitly included as a facet in both models. This decision was based on prior findings that replicate evaluations enhance measurement reliability and are essential for monitoring assessor performance.

TIM: $ln\left(P_{mnrik}/P_{mnrik-1}\right) = \beta_m - \theta_n - \rho_r - \delta_i - \tau_k$ *...Equation 4.1*

DFCM: $ln\left(P_{mnrk}/P_{mnrk-1}\right) = \beta_m - \theta_n - \rho_r - \tau_k$ *...Equation 4.3*

Where: in the DFC models (DFCM), the δ_i parameter was not included due to the absence of attributes in the analysis.

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

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

β_m= degree of leniency or severity of assessor (m) in rating attribute intensities

θ_n= degree of difference in the total intensity measure for sample (n)

ρ_r = degree of difference between ratings of samples in replicate session (r)

δ_i = the average degree of intensity of sensory attribute (i) across all samples

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

6.2.6.1 Selection of assessors based on model fit (TIM)

A subset of assessors was identified based on their TIM model fit statistics. Assessors whose OUTFIT Mnsq values fell within the acceptable fit range (discussed later in section **6.3.5**), were classified as *Selected assessors*. A separate TIM model was then fitted using only this subset to determine whether panel discrimination and diagnostic clarity improved compared with the full panel.

6.2.6.2 Statistical analyses

Statistical analyses were conducted for the DFC raw scores and all Rasch models (including the DFC Rasch measures and TIM). The TIM analysis was performed for both the full set of assessors and the selected assessors, and the results were compared in terms of discriminatory ability and diagnostic detail.

6.3 Results and Discussion

6.3.1 Fit of data to the Many-Facet Rasch Model (MFRM)

The results of the global model fit, *Assessor facet* fit statistics and response dependency checks are presented in **Table 6.2**.

Table 6.2 Summary of Rasch model fit statistics and response dependency results for DFC and TIM models.

Criteria / Assessors	TIM		DFCM
	All (n=54)	**Selected** (n=17)	**All** (n=54)
Global fit (*StRes*)[1]			
≤5% ≥ 2	4.6 (403)	4.1 (113)	3.7 (18)
≤1% ≥ 3	0.2 (20)	0.0 (0)	0.6 (3)
Total	8748	2754	486
Assessor Fit			
OUTFIT Mnsq*[2] *(N_r =162)			
% Fit (0.78-1.22)	31 (17)	100 (17)	—[3]
%Overfit (≤ 0.78)	39 (21)	0 (0)	—
% Underfit (≥1.22)	28 (15)	0 (0)	—
%Extreme Misfit (>2.0)	2 (1)	0 (0)	—
Unidimensionality[4]			
1st contrast eigenvalue (<2)	2.45	2.69	—
LID (attributes)[5]			
Corr. of StRes (<0.3)			
Rich Aroma - Savoury Flavour	NA[6]	0.48	—
Smooth Mouthfeel - Homogenous Mouthfeel	0.47	0.45	—
Viscous Appearance - Thick Mouthfeel	0.39	0.42	—
Pungent Aroma - Rich Aroma	NA	0.30	—
Herby Appearance - Viscous Appearance	NA	0.28	—
Sour Taste - Salty Taste	0.25	NA	—
Cooked Tomato Aroma - Rich Aroma	0.24	NA	—

[1] Percentage (number of observations in brackets) of absolute standardised residuals (StRes).

[2] Outlier-sensitive measure of unweighted mean squares indicating deviation of the *Assessor facet* estimates from Rasch model predictions. The acceptable fit range (0.78-1.22).

[3] — indicates that assessor-performance diagnostics and response-dependency checks were not applicable for the DFCM, as this model was used only to compare overall difference results with the TIM-derived measure.

[4] Eigenvalue of the unexplained variance in the first contrast, not accounted for by the Rasch model, in PCAR.

[5] Local Item Dependency (LID) examined through the correlation of standardised residuals (Corr. of StRes) between attributes, with values > 0.3 indicating that items (attributes) are dependent.

[6] NA =Not applicable meaning attributes were not flagged as potentially dependent for the panel.

To recap, an acceptable global model fit of the data is when about 5% or less of absolute standardised residuals is ≥2, and about 1% or less is ≥3 (Linacre, 2022; Eckes, 2023). Both the TIM and DFCM models showed an acceptable global fit, suggesting that overall, the data in each model aligns with the assumptions of the Rasch model, with no major inconsistencies likely to distort measurement.

Only the *Assessor facet* fit statistics for the TIM models (All and Selected assessors) are presented, as all other facets demonstrated 100% fit in both models. Assessor performance was not evaluated for the DFCM because this model was used solely to compare the overall difference measure with that obtained from TIM. Therefore, assessor-level diagnostics apply only to the TIM models. The DFCM includes only one item (DFC), so attribute response-dependency checks are also not applicable.

The acceptable OUTFIT Mnsq range for assessors was calculated* as 0.78 - 1.22, based on a total of 162 responses per assessor. The results showed that more than half of the assessors in the full panel, exhibited response patterns that deviated from the expectations of the Rasch model. As a result, a subset of assessors with OUTFIT Mnsq values within the acceptable range was selected, and their data were fitted to a separate Rasch model (labelled *Selected* on **Table 6.2**), which showed a 100 percent fit for the *Assessor facet*.

Unidimensionality and Local Item Dependence (LID) for the TIM model were examined by Principal Component Analysis of Residuals (PCAR). Unidimensionality is confirmed when the eigenvalue of the unexplained variance in the first contrast is <2, and Local item dependence (LID) is identified when the residual correlation between two attributes exceeds 0.3 (Ramp et al., 2009; Christensen et al., 2017).

The results in **Table 6.2**. reveal that for both the full and selected panels, the unexplained variance in the 1st contrast of the Rasch Principal Component Analysis of Residuals (PCAR) had eigenvalues of 2.45 and 2.69, respectively. This might suggest the presence of a minor secondary dimension, roughly equivalent to the strength of 2-3 items.

* $1 \pm 2\sqrt{\frac{2}{Nr}}$ (Wu & Adams, 2013; Eckes, 2023), where Nr (number of responses) for each assessor is 162.

Examination of standardised residual correlations revealed that several pairs of attributes were locally dependent, meaning that their co-variation exceeded what would be expected by chance (Linacre, 2024a). Most correlations above the conventional 0.3 threshold indicate systematic relationships that warrant further investigation.

Specifically, four attribute pairs exhibited strong residual correlations: *Smooth Mouthfeel* and *Homogeneous Mouthfeel* (0.47 and 0.45), and *Viscous Appearance* and *Thick Mouthfeel* (0.39 and 0.42), for the full and selected panels, respectively. In the selected panel, additional dependencies emerged between *Rich Aroma* and *Savoury Flavour* (0.48), and *Pungent Aroma* and *Rich Aroma* (0.30). These dependencies are not considered problematic, especially since the latent variable being measured is the overall difference between samples. Instead, they provide useful information about how certain attributes tend to vary together across samples. These patterns reflect genuine differences among the samples, rather than distortions in the measurement model and may arise from attributes assessing the same physical property through different senses (e.g., viscosity) or from cross-modal sensory interaction (e.g., aroma-flavour perception).

For example, if a tomato soup sample is rated as having a more *Viscous Appearance*, it is also likely to be rated as having a *Thick Mouthfeel*. The same pattern is seen between Rich Aroma and Savoury Flavour, and between *Smooth Mouthfeel* and *Homogeneous Mouthfeel*. Notably, the selected assessors seemed more aware of the connection between *Rich Aroma* and *Savoury Flavour*, suggesting a higher level of sensitivity or consistency.

Attribute pairs below the 0.3 threshold, while not indicating local dependency, still provide useful diagnostic insights. *Herby Appearance* and *Viscous Appearance* showed a correlation of 0.28 for the selected assessors, indicating a potential link between these attributes. This may be explained by the sample preparation (see **Table B 3**). Sample A was prepared using a cream of tomato soup base, with extra cream, and dried chopped basil, added during heating, to align more closely with the other samples, which were made from a cream of tomato and basil soup. However, because the basil in Sample A was added only while heating shortly before serving, it was likely less integrated into the soup, compared to the basil in the other

samples, which had been incorporated during manufacturing and packaged for shelf stability. This may have affected the appearance of Sample A, contributing to higher ratings for *Herby Appearance*, and likely explains its flagged correlation with *Viscous Appearance*.

Similarly, for the full panel, weak correlations were found between *Sour Taste* and *Salty Taste* (0.25), and between *Cooked Tomato Aroma* and *Rich Aroma* (0.24). Although these values were below the dependency threshold, they reflect known sensory interactions. For instance, Fabian and Blum (1943) found that sodium chloride (NaCl), at sub-taste threshold levels, reduced sourness in foods, while Breslin (1996) reported that acids, (such as the citric acid found in tomatoes), enhanced saltiness perception. More recent work has shown that salt and sour mixtures can mutually enhance each other at low intensities and show suppression or no effect at higher intensities (Keast & Breslin, 2002; Liem et al., 2011). Additionally, the complexity of tomato soup aroma, influenced by various volatile compounds(Kazeniac & Hall, 2006; Gilsenan, 2010; Distefano et al., 2022) likely contributed to the perception of rich aroma across samples. This was likely due to Sample B, which contained added passata and showed higher *Rich Aroma* ratings than the other samples (**Table E 2**).

6.3.2 Rating scale category diagnostics

Following the established guidelines outlined in **Table 3.1**, deviations in the interpretation and operational use of the scale, relative to the expectations of the Rasch model, were empirically investigated.

The category functioning of the rating scales for the attributes rating (AR) intensity and DFC rating scales are presented in **Table 6.3**. All essential criteria for measure accuracy and for description of the tomato soup samples in the study were met. Specifically, the Rasch-Andrich thresholds were ordered, and no misfitting categories were observed, as OUTFIT Mnsq values were close to 1.0. This suggests that responses to attributes in the models, are consistent with estimates of the latent variable (Tennant & Conaghan, 2007), and meet the model expectations. Additionally, the observed average measures increased monotonically across the scale categories, indicating that no scale categories were skipped along the variable (Eckes, 2023).

Table 6.3 Summary of scale category statistics for Intensity and DFC rating scales used in the TIM and DFCM models

Panel	Scale Categories		Frequency[1]	Average Measure[2]		OUTFIT Mnsq[3]	Rasch Andrich Threshold	
				Observed	Expected		Measure	Distance[4]
TIM All Intensity Rating Scale 8-category 01234567	0	None	149 (2)	-0.62	-0.60	1.0		
	1	Barely detectable	539 (6)	-0.47	-0.50	1.1	-1.84	**0.30^**
	2	Weak	1614 (18)	-0.37	-0.38	1.0	-1.54	0.80
	3	Moderate	2452 (28)	-0.26	-0.25	0.9	-0.74	0.62
	4	Strong	2305 (26)	-0.13	-0.11	1.0	-0.12	0.72
	5	Very strong	1231 (14)	0.07	0.05	1.0	0.60	0.64
	6	Extremely strong	408 5)	0.25	0.22	1.0	1.24	1.17
	7	Strongest imaginable oral sensation	50 (1)	0.36	0.40	1.0	2.41	
TIM Selected Intensity Rating Scale 8-category 01234567	0	None	27 (1)	-0.76	-0.71	0.9		
	1	Barely detectable	200 (7)	-0.64	-0.61	0.9	-2.66	1.17
	2	Weak	505 (18)	-0.48	-0.50	1.1	-1.49	0.69
	3	Moderate	718 (26)	-0.38	-0.39	1.0	-0.80	**0.42^**
	4	Strong	763 (28)	-0.28	-0.26	1.0	-0.38	0.72
	5	Very strong	447 (16)	-0.11	-0.13	1.0	0.34	1.18
	6	Extremely strong	91 (3)	-0.06	0.00	1.1	1.52	1.95
	7	Strongest imaginable oral sensation	**3 (0)***	0.38	0.12	0.9	3.47	
DFCM All DFC Rating Scale 7-category 0123456	0	No difference	65 (13)	-1.20	-1.34	1.2		
	1	Barely detectable difference	70 (14)	-0.94	-0.91	1.0	-1.21	**0.47^**
	2	Slight difference	75 (15)	-0.50	-0.42	1.1	-0.74	**0.36^**
	3	Moderate difference	91 (19)	-0.04	0.04	1.2	-0.38	0.59
	4	Large difference	91 (19)	0.32	0.36	1.0	0.21	0.74
	5	Very large difference	57 (12)	0.69	0.60	0.7	0.95	**0.21^**
	6	Extremely different	37 (8)	0.98	0.86	0.9	1.16	

[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 Mnsq 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 thresholds refers to the spacing between adjacent response categories, which is 0.51 and 0.57 for 8 and 7-point category scales, respectively.

^ Insufficient minimum advancing distance between Rasch-Andrich thresholds suggesting that adjacent categories are less distinctive than intended (helpful for inference on subsequent studies).

* Each scale category should have at least 10 observations as this is essential for measure stability.

A criterion for ensuring measure stability (**Table 3.1**) was not met for the selected assessors. Measure stability refers to the consistency of a measurement system when repeated over time in the same context. Specifically, there were only 3 observations in the highest category of the Intensity scale (7 = Strongest imaginable oral sensation), whereas Linacre (2002b) recommends a minimum of 10 observations per category. Additionally, all models failed to meet the minimum advancing distance between category thresholds at least once, indicating that some categories were too close together and less distinctive than intended (Eckes, 2023). The recommended minimum distance is 0.51 and 0.57 for 8 and 7-point scales (Ho, 2019).

For the TIM full panel, this issue occurred between categories 1 and 2 (barely detectable and weak, respectively), while for the TIM selected panel, it was between categories 3 and 4 (moderate and strong)The issue was more pronounced in the DFCM, effectively showing that only four categories were clearly distinctive: 0- no difference, 3- moderate difference, 4- large difference, and 5- very large difference, with category 4 nearly failing to meet the threshold at 0.59. This suggests that assessors may have had difficulty distinguishing between adjacent categories possibly due to overlapping interpretations of the scale category descriptors.

This insight is valuable for developing rating scales in sensory quality programs for specific products, as meeting these criteria improves the reliability of inferences in future studies. The recommended remedial action is to collapse the affected adjacent categories before data collection in subsequent studies, provided the panel is sufficiently consistent. However, no changes to the scale were made in this study because scale development was beyond the scope.

Conventionally, the rating scales used for attribute evaluation in descriptive analysis like the QDA are relative scales anchored to the attribute of interest, often established using reference samples (Meilgaard et al., 2025), rather than the absolute end-anchors used in the LMS scale (adapted for this study). Absolute anchors such as “strongest imaginable oral sensation” are interpreted against a much broader experiential frame of reference, so assessors may reserve the highest category for exceptionally unusual sensations (Lawless & Heymann, 2010). Therefore, the limited use of the upper-end categories in this study was not

surprising and reflects how the MFRM diagnoses the operational use of the scale based on the observed response patterns. An interesting direction for future research would be to examine how the MFRM performs with the relative scales commonly used in descriptive analysis.

6.3.3 Representing the Overall Difference Construct

Wright maps for TIM models for the full and selected panels, as well as the DFCM model, are presented in **Figure 6.1**, **Figure 6.2**, and **Figure 6.3** respectively, with all four facets (*Assessors*, *Samples*, *Repetition*, and *Attributes*) positively oriented, as described in previous chapters. The *Sample facet* was non-centred, while the other facets were centred at the mean (0 on the logit scale) to serve as a reference point. Consequently, sample locations were adjusted by considering the severity of assessors, the average intensity of attributes, and the intensity ratings in repeated sessions, representing the *Assessor, Attribute*, and *Repetition facets*, respectively. In the *Assessor facet*, assessors with higher logit values are more lenient, generally assigning higher scores on the rating scale; in the *Sample facet*, samples with higher logit values have higher Total Intensity Measure (TIM) or, for the DFC measure (DFCM), are more different from the control; in the *Repetition facet*, replicate sessions where higher intensity ratings were assigned on average have higher logit values; and in the *Attribute facet*, attributes with higher average intensity ratings have higher logit values.

6.3.3.1 Total Intensity Measure Representation (Full and Selected Panel)

Figure 6.1 presents an overview of the full panel's ratings of the overall difference between the tomato soup samples. Assessors' severity estimates were distributed around the mean within a range of approximately -0.6 to 1.0 logits (S.E = 0.07), indicating meaningful differences in their severity of scale use, after accounting for measurement error. Assessors 3029 and 3016 emerged as the most lenient, as they were positioned noticeably higher than the rest of the panel on the map.

The average attribute intensity ratings for each sample were below the mean (0 on the logit scale), and differences across the three replicated sessions were not significant. Samples positioned higher on the scale were perceived as having greater intensity, based on ratings averaged across all attributes.

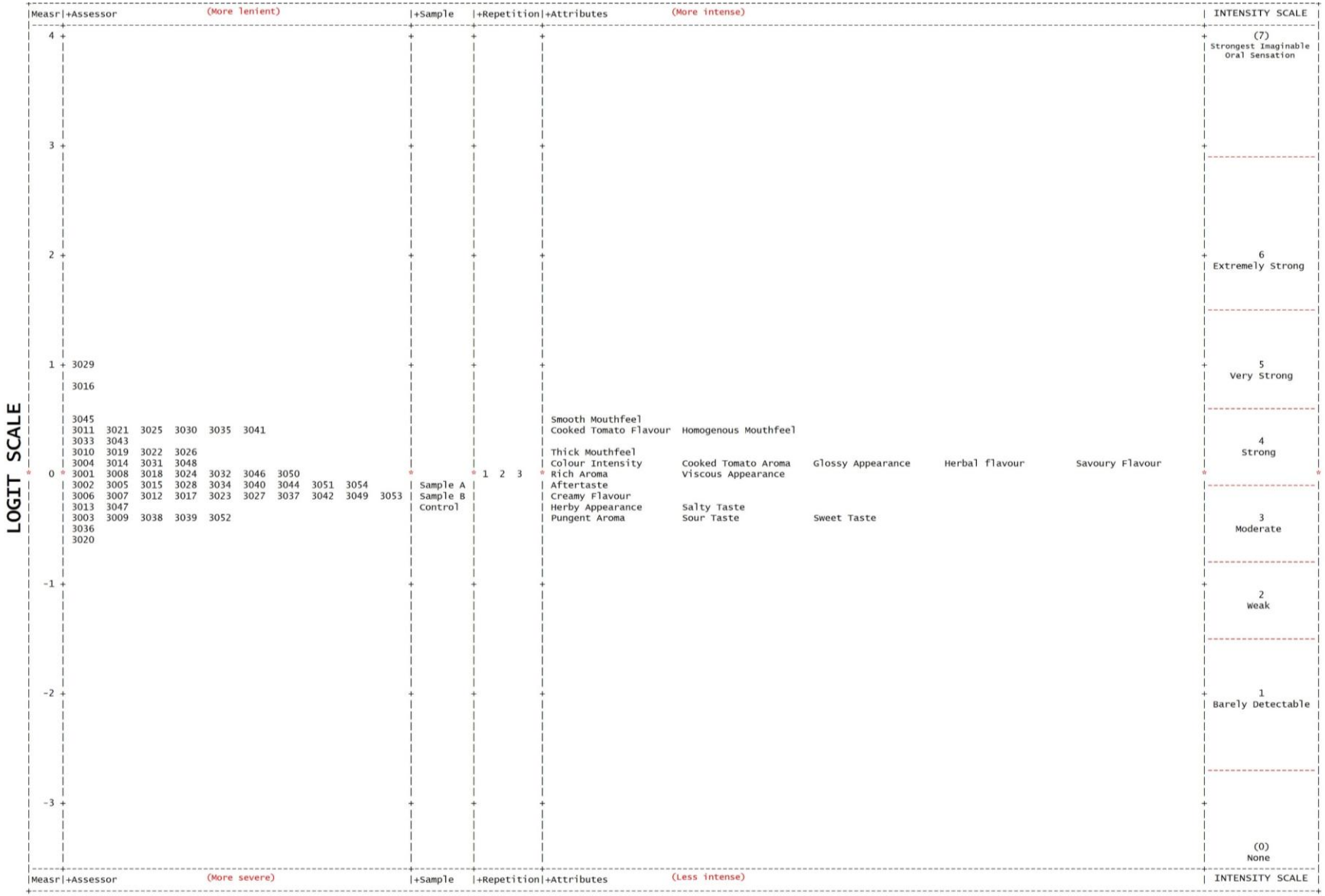


Figure 6.1. Wright map for the TIM model representing All Assessors (Assessor IDs 3001 - 3054)

The latent variable of Overall Difference is reflected in the Total Intensity Measure (TIM), shown by each sample's position on the logit scale. Although the three samples were separated by only 0.3 logits (S.E = 0.02), Rasch separation statistics will determine whether these differences are statistically meaningful. The TIM values will also be used in pairwise comparison tests to evaluate the extent of differences between each sample and the control (discussed in section **6.3.4**).

From the *Attribute facet* and intensity scale, most attributes were, on average, rated between categories 3-Moderate intensity and category 4-Strong intensity (see **Table E 2**). Mouthfeel attributes emerged as the most intense, while taste attributes were the least intense. Among them, Smooth mouthfeel was the most dominant attribute perceived by the panel across all samples. The OUTFIT Mnsq statistics for individual attributes (discussed later in the chapter) will indicate which attributes are primarily driving differences between the samples.

In terms of scale category usage, the intensity scale on the Wright map showed that attribute ratings were evenly distributed from category 2-Weak to category 5-Very strong. The increasing width of half-point thresholds beyond these points on both ends suggests less frequent use of the extreme categories. This pattern is also reflected in the scale category statistics presented in **Table 6.3**.

For the selected assessors presented in **Figure 6.2**, the assessors were more tightly clustered around the mean, within a range of approximately −0.4 and 0.4 logits (S.E = 0.07). This suggests smaller differences in scale use severity across the assessors after accounting for measurement error. However, the range remains substantial, and any meaningful differences will be confirmed through separation statistics discussed later.

As with the full panel, the average attribute intensity ratings across all samples remained below the mean (0 on the logit scale). Samples positioned higher were perceived to have greater intensity, based on ratings averaged across all attributes.

In this panel, the distinction between the two test samples and the control sample was more pronounced than in the full panel. In the *Repetition facet*, differences across the three replicated sessions showed slight variation, they may not be statistically significant. However, Ratings were generally lower in the first replicate,

likely reflecting initial uncertainty. By the second and third replicates, ratings had stabilised as assessors developed a clearer mental frame of reference.

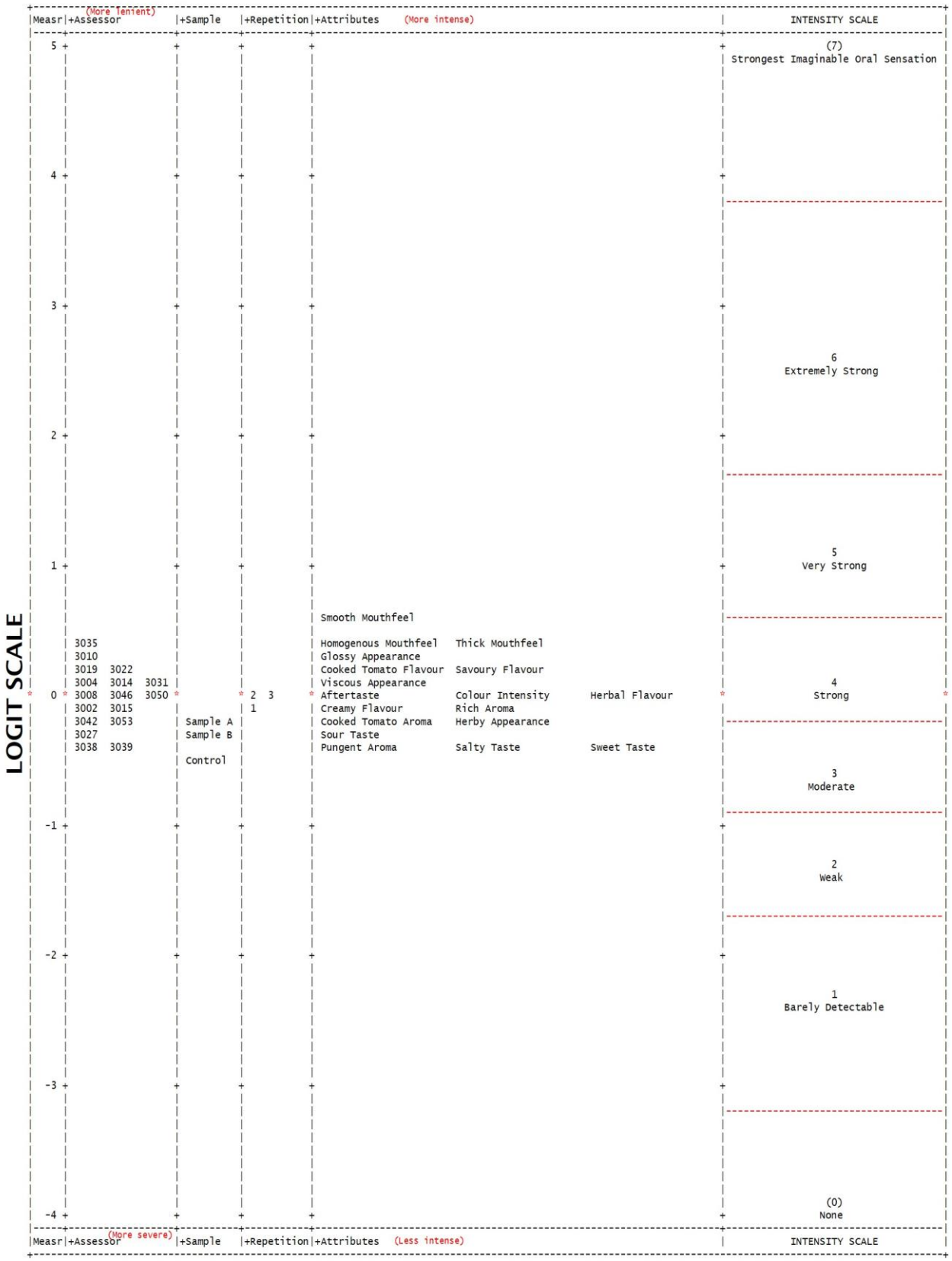


Figure 6.2. TIM model Wright map representing the 17 Selected Assessors (IDs 3002 - 3052)

The latent variable of overall difference is captured in the Total Intensity Measure (TIM), represented by the position of each sample on the logit scale. This indicated that Sample A and Sample B were less different from each other than either of them was from the control sample. Rasch separation statistics will determine whether these differences are statistically significant, and TIM values will be used for pairwise comparisons with the control.

Like the full panel, the *Attribute facet* and intensity scale showed that mouthfeel attributes were rated most intensely, with Smooth mouthfeel being most dominant across samples, while taste attributes consistently received the lowest intensity ratings. Scale category usage patterns were also comparable to those observed in the full panel.

6.3.3.2 DFCM representation for the full panel

Figure 6.3 shows the Wright map for the Difference from Control measures (DFCM) for the full panel of assessors. The *Assessor facet* reveals varying degrees of severity in scale use, with assessors distributed across a range from approximately -1.5 to 1.7 logits (S.E = 0.27- 0.36) around the mean. Assessor 3054 was the most severe, with a 0.7 logit gap from the next closest assessor. The distribution also indicated several distinct levels of severity among assessors, particularly toward the more severe end of the scale.

As shown in the Wright map for the selected assessors (**Figure 6.2**), Sample A and Sample B were less different from each other than either of them was from the control sample. The control sample was clearly different, positioned approximately 1.5 and 1.6 logits away from Sample B and Sample A respectively. While there appear to be slight variations across replicate evaluations, these differences may not be statistically significant. The significance of these differences will be further examined using Rasch separation statistics and pairwise comparison tests.

It is important to note that on the Wright maps for the DFC model, the control sample values reflect the ratings assigned to the blind control, and the map effectively represents the magnitude and direction of perceived differences between the control and test samples. Unlike the findings in Chapter 4, where the

control appeared close to one of the test samples, the current results show that assessors more accurately rated the blind control as identical to the actual control.

This improvement is evident in both the DFC Wright map and the category diagnostics (**Table 6.3**), where a sufficient number of assessors assigned the blind control a rating of 0, indicating no perceived difference from the control sample. These results suggest greater consistency in identifying the control sample in the present study, possibly due to more noticeable compositional differences between the control and the test samples.

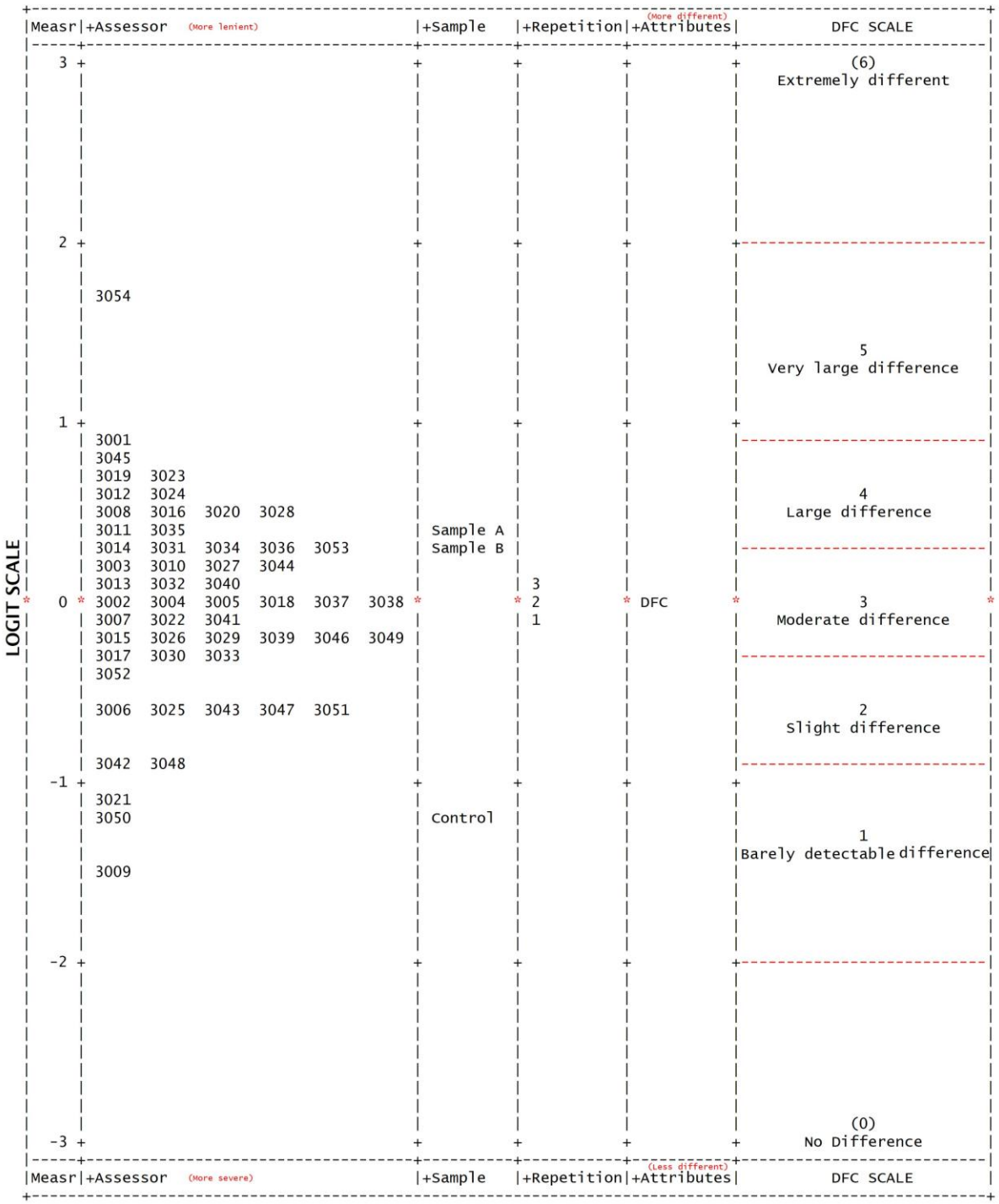


Figure 6.3. DFCM Wright map representing All Assessors (Assessor IDs 3001 - 3054)

However, ratings were evenly distributed across the DFC scale, suggesting some degree of imprecision in the ratings assigned by the assessors on average. In a proper-functioning scale, some categories would be used more than others to reflect actual differences in performance. An even spread suggests that assessors were not consistently distinguishing subtle differences in performance, or that some categories overlap in meaning. As noted by Bond et al. (2020) and Eckes (2023), intermittent low-frequency categories indicate irregular scale usage and the presence of redundant categories. Nonetheless, the improved identification of the blind control, highlights the value of targeted assessor preparation in enhancing rating accuracy.

A comparison of the Wright maps for both the full panels (**Figure 6.1** and **Figure 6.3**) and the selected TIM panel (**Figure 6.2**), revealed a consistent pattern in the relative positioning of the test samples, compared to the control. Sample A was located higher than Sample B, and Sample B higher than the Control sample. While this pattern was not visually apparent in the TIM model with all the assessors, further analysis using Rasch separation statistics and pairwise comparisons will provide additional insights.

6.3.4 Rasch separation statistics, panel performance, and comparison of overall sample differences

The Rasch separation statistics for assessors, repetitions, and samples, along with the pairwise comparisons for the samples, are presented in **Table 6.4**.

The fixed Chi-square (χ^2) Rasch separation statistic tests the null hypothesis that no meaningful differences exist within a given facet. For the assessors, it serves as an indicator of panel agreement in the use of the scale, testing whether, after accounting for measurement errors, (such as inconsistent ratings across replicate evaluations or differing rank ordering of samples), the severity of all assessors is the same (Myford & Wolfe, 2004; Linacre, 2025c).

The χ^2 values for all the panels, that is, the full and selected TIM panels as well as the DFCM panel, revealed that assessors were not in agreement, as all panels showed highly significant χ^2 values ($p<0.001$). A significant χ^2 indicates that the variation in the severity levels of the assessors is greater than would be expected by

chance, meaning that, in each panel, at least some assessors were systematically more severe or more lenient than others in the ratings they assigned.

Strata values, which indicate how many statistically distinct levels of severity exist among the assessors, showed that there were approximately 7, 5, and 3 distinct groups of assessors in the TIM full, TIM selected and DFCM models respectively. However, the reliability of separation, which reflects how precisely these differences in severity are measured relative to the error in the estimates, showed values of 0.96, 0.91, and 0.75 respectively for the TIM full, TIM selected, and DFCM panels. These results suggest that differences in the DFCM panel were somewhat influenced by measurement errors.

Table 6.4 Summary of Rasch separation statistics and sample comparisons

		TIM		DFC (All Assessors)	
Test/Dataset[1][2]		All	Selected	DFCM	DFC Raw
Rasch Separation Statistics					
Assessors	Reliability $_{Assessor}$	0.96	0.91	0.75	
	Strata $_{Assessor}$	6.50	4.52	2.63	
	Fixed χ^2	1169.5***	182.5***	182.1***	
Repetition	Reliability $_{Repetition}$	0.44	0.80	0.11	
	Strata $_{Repetition}$	1.52	2.98	0.80	
	Fixed χ^2	5.4$^{~}$	14.9***	3.4	
Samples	Reliability $_{Sample}$	0.97	0.89	0.99	
	Strata $_{Sample}$	8.16	4.14	14.89	
	Fixed χ^2	**106.3*****	**27.5*****	**325.9*****	
Pairwise Comparisons					
Kruskal-Wallis Test (H)[3]		**406.15*****	**124.4*****	**286.89*****	**180.92*****
Mean differences (Dunn's Many-to-One Test)					
Control-Sample A		-0.22***	-0.20***	-1.61***	-2.39***
Control-Sample B		-0.09***	-0.14***	-1.50***	-2.22***

[1] P-value levels of significance: <0.001***, <0.01**, <0.05*, <0.1$^{~}$; measures with no superscript symbols have p-values >0.1.

[2] For degrees of freedom (df) = 2, the chi square (χ^2) critical values are 5.991 ($\alpha = 0.05$) and 4.605 ($\alpha = 0.1$).

[3] The Kruskal-Wallis test statistic (H) also follows a χ^2 distribution for determining significance

Reliability values closer to 0 for assessors are desirable, as they suggest that there is no statistical distinction between lenient and severe assessors (Myford & Wolfe, 2004). However, as noted by Wright and Masters (1982) and, by Bond et al. (2020) lower reliability, when significant differences are observed, also implies some degree of imprecision in the estimates, possibly due to inconsistent use of the scale categories, as seen in the DFCM Wright map and scale category statistics.

Separation statistics for the *Repetition facet* (**Table 6.4**), supported the presence of a meaningful difference between the replicate evaluations of the selected assessor panel in the TIM model, with a highly significant χ^2 ($p<0.001$), a strata of 2.98 and reliability value of 0.80, as was also suggested by the Wright map illustration. In contrast, the lower Strata value of 1.52 and a barely significant χ^2 ($p<0.10$) for the full panel, along with a low reliability value of 0.44, indicated that any observed differences were likely due to measurement error. This shows that the full panel produced more consistent results across replicate evaluations, as lower strata and reliability in this facet indicate that there were no meaningful differences between replicate sessions.

In the DFCM panel, however, no meaningful differences were found between replicate evaluations. This may reflect the design of the DFC presentation, which requires assessors to make a direct comparison against a physical control sample. This is cognitively simpler and may help stabilise judgements across repetitions. In contrast, the monadic presentation used in the attribute rating test requires assessors to develop and refine an internal frame of reference over time, which can introduce greater variability across replicates. A similar pattern was observed in Chapter 4: **pg.94**, where the DFC results similarly showed more stable replicate evaluations and greater overall discrimination than the TIM results.

For the *Sample facet*, separation statistics assess the panel's ability to distinguish meaningfully between the samples. Higher reliability values, ideally close to 1, indicate that the differences observed between samples are consistent and not due to measurement error. In this analysis, all panels demonstrated statistically significant differences between samples ($p<0.001$). The Strata values suggest that the full TIM panel and the DFCM panel could differentiate approximately 8 and 15 statistically distinct levels among the sample measures, respectively. These high

values imply considerable variation within the samples, likely reflecting differences among assessors and/or measured attributes. For the TIM full panel, this differentiation is likely influenced by both the diversity of the 18 attributes and inconsistencies among assessors.

In contrast, the DFCM panel, comprising only three samples and a single item in the *Attribute facet*, achieved an even higher Strata value (15). The comparative test design was likely the major contributor. As discussed in Chapter 4: **pg.89** the DFC test design involves a direct comparison with a control using a difference scale anchored to the perceived intensity of that control. This relative judgement is cognitively simpler because the control provides a stable reference point for every evaluation. Consequently, less cognitive effort is required to rate the test sample than in the attribute-rating tests, where samples were presented monadically, one at a time, and assessors were required to rate multiple attributes using absolute scales anchored with descriptors such as *strongest imaginable sensation*. Without an external comparison reference, assessors must rely on their own mental reference, which is more susceptible to inconsistency, introducing greater variability into the ratings. Additionally, the higher Strata in DFCM compared to TIM likely reflects that DFCM directly measured overall difference (assessors rated "difference from control"), while TIM derived it as a latent variable from 18 separate attribute intensity ratings. This difference between direct measurement and derived estimation may explain the observed separation differences.

Individual differences in how assessors applied the scale may also have inflated the separation. This interpretation is supported by the Wright map (**Figure 6.3**), which shows that the spread of assessor severity was wider than the spread of sample measures, and by the scale category statistics (**Table 6.3**), which indicate that ratings were widely and inconsistently distributed across the scale categories.

Meanwhile, the selected TIM panel produced a more conservative Strata value of 4.14, indicating that about 4 distinct levels could be reliably identified. This suggests that the variation in ratings was more aligned with expected sample difference, given that there were only 3 samples, with less additional variation attributable to assessors or other factors. High reliability values across all facets support that these distinctions are meaningful and not due to random error.

Notably, the selected TIM panel showed a lower sample separation reliability (0.89) compared to the full TIM panel (0.97) and DFCM panel (0.99), a reduction that may be explained by inconsistencies across replicate evaluations, as indicated by the significant χ^2 ($p<0.001$) in its *Repetition facet*.

Pairwise comparisons were conducted using the non-parametric Kruskal-Wallis test on the Rasch measures, which has been shown to be robust for such data as discussed in Chapter 4. Although parametric tests were initially performed, several parametric assumptions, including normality were violated, consistent with findings from the other Rasch measures in this study. Previous research has shown that Rasch measures of latent variables violates normality assumptions (Guilleux A, 2014; Ho, 2019; Lacko, 2023). However, non-parametric methods provide a robust alternative and are well-suited for Rasch measures as discussed in Chapter 4: **pg.92**. As shown in **Table E 1**, sample comparison results from the parametric Tukey's HSD test were generally consistent with those from the non-parametric Kruskal-Wallis test, suggesting that the non-parametric approach does not compromise the validity of the findings. Furthermore, the Kruskal-Wallis test results were more closely aligned with those from the Rasch fixed chi-square (χ^2) test.

The mean difference results in **Table 6.4** show that both Sample A and Sample B were significantly different from the control sample, with the direction of these differences being relatively consistent across the DFC and Rasch-based TIM approaches, as well as across the panels. Based on the sample formulations (**Table B 3**), these differences were expected; however, it was unclear whether differences would be more pronounced in textural properties or flavour across the samples. The results indicate that Sample A, which was designed to be thicker than the other samples due to the addition of double cream, exhibited a greater magnitude of difference from the control compared to Sample B, which was engineered to have a stronger savoury flavour, due to the addition of garlic granules.

A review by Tournier et al. (2007) shows that changes in rheological properties, such as viscosity in the case of Sample A, can affect the perception of aroma (Ferry et al., 2006; Lubbers et al., 2007), flavour, and taste (Hollowood et al., 2002; Saint-Eve et al., 2004) attributes through cross-modal interactions. In a recent study on beef broths enriched with taste enhancers, Brouwer et al. (2024) found that increasing

viscosity of beef broths enriched with sodium chloride, MSG and Kokumi compounds resulted in more intense and richer savoury, salty and beef flavour. More broadly, Wang et al. (2025) reported that components of the food matrix, including proteins, lipids and carbohydrates, can bind or entrap aroma compounds and alter their release. This indicates that viscosity and matrix structure can substantially influence how aroma and flavour are perceived in real foods. These sensory interactions may have contributed to the increased perceived complexity of Sample A beyond what would be expected from a typical cream of tomato and basil soup.

Additionally, the panel perceived mouthfeel attributes as the most dominant, and Sample A likely had a noticeably different mouthfeel compared to the other samples. Mouthfeel attributes are often more readily perceivable compared to other sensory attributes, due to their somatosensory and tactile nature (Lawless & Heymann, 2010; Stone et al., 2012; Ditschun et al., 2025).

These findings suggest that textural modifications, such as increased viscosity in Sample A, not only produce larger perceived differences but may also have enhanced flavour complexity through multisensory interactions. Furthermore, mouthfeel attributes, being more directly perceivable, contributed to these pronounced differences. Overall, the TIM method was sensitive to subtle sample variations across the combined attributes and provided more diagnostic insight than the broader DFC tests, highlighting its advantages for evaluating targeted sensory attributes in difference testing.

To address a limitation identified in the previous study, specifically the non-representative choice of attributes in the AR test, a comment section was added to the end of the DFC test questionnaire. Assessors were asked if they had considered any additional attributes beyond those listed when evaluating differences between the samples. The responses showed that either no new attributes were identified or that any additional descriptors mentioned, corresponded to terms or ingredients already captured in the attribute list (**Table 6.1**), with assessors generally commenting on the most prominent characteristics. Examples of their comments include:

"... creamy textures were more obvious"

"...how the soup coated the entire tongue / mouth"

"Two of the test samples had a really nice cheesy note..."

"...one had more garlic (garlic granules taste) the other was thicker (maybe xanthan)"

"...one definitely had beef stock in it"

Since the comments aligned with the existing attribute descriptors in the lists, it is reasonable to conclude that the AR tests in this study effectively captured the perceivable differences observed in the DFC test.

The next section examines the specific attributes perceived as driving the most differences, as well as assessor rating behaviour.

6.3.5 Individual assessor performance analysis

OUTFIT Mnsq values for the *Assessor facet* were used to evaluate rating behaviour in the TIM model. The trends and results observed across the assessors were consistent with those reported in section **5.3.4**, where certain value ranges corresponded to specific rating behaviours. However, as previously noted, the Rasch approach to evaluating individual assessor performance is always relative to the performance of other assessors in the analysis (Myford & Wolfe, 2004). Therefore, it is most informative when applied to a more homogeneous panel, as demonstrated in Chapter 5 with the trained panel. In contrast, the untrained panel in Chapter 5 showed greater variability, which limited the usefulness of the diagnostics. Since the assessors in the current study were also untrained, similar variability is likely.

Considering this, the analysis in this chapter focuses on identifying assessor OUTFIT Mnsq ranges (**Figure 6.4**) and their implications for rating effects. The subset of assessors discussed so far was selected based solely on these metrics, unlike in the previous chapter, where selection was informed by both OUTFIT Mnsq insights and the assessors' discriminatory ability, as determined by individual ANOVAs.

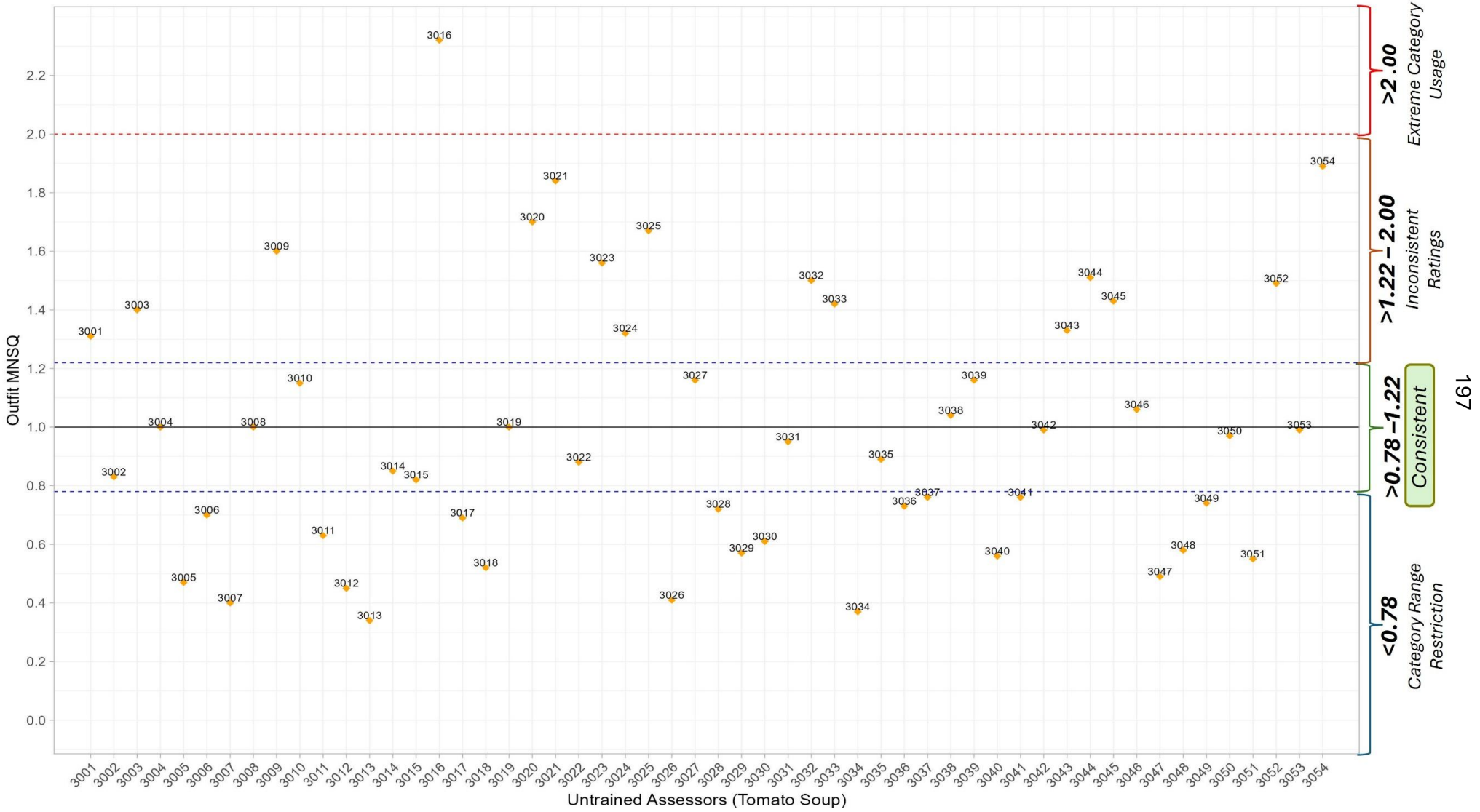


Figure 6.4. OUTFIT Mnsq plot for the TIM full untrained tomato soup panel

In the discussions (section 5.3.4.3.2: **pg.154**), it was hypothesised that, just as greater variance among parameters within a facet leads to higher OUTFIT Mnsq values (Linacre, 1995), lower variance corresponds to lower values. This relationship may reflect the severity of overfit rating effects, such as restriction of range and central tendency. Each rating effect spans a continuum, and OUTFIT Mnsq values vary depending on how strongly an assessor exhibits a particular rating behaviour. These patterns are examined in this section.

Figure 6.4 above presents the OUTFIT Mnsq control chart for all assessors in the TIM model. The acceptable OUTFIT Mnsq range for assessors in this study was 0.78 to 1.22. Assessors with values below the lower limit are considered overfit, exhibiting range restriction and central tendency rating effects. Those with values above the upper limit are underfit, showing more erratic ratings either within themselves or compared to the rest of the panel. Additionally, OUTFIT Mnsq values above 2.0 indicate that an assessor is using extreme ends of the rating scale, which can skew the overall panel results and cause misleading conclusions.

The subset of assessors discussed earlier were those with values strictly within the acceptable range of 0.78 to 1.22. These assessors, identified for their relatively consistent ratings, will be used in subsequent analyses to determine key discriminating attributes across the samples.

The following **Figure 6.5,** spanning across two pages, presents response distribution (trellis) plots for assessors whose OUTFIT Mnsq values fall below the lower threshold of 0.78, indicating potential overfitting. This illustration supports the hypothesis that rating effects are reflected by the OUTFIT Mnsq values along a spectrum. From top to bottom, each row represents an assessor's responses across several attributes. Assessors are arranged from the lowest OUTFIT Mnsq value (most overfitting) up to the 0.78 threshold. The corresponding OUTFIT Mnsq values are displayed in the rightmost column. The black dotted line marks the margin for the central scale categories, specifically scale usage between 2 and 4, based on the most frequently used categories across the panel (as shown in the scale category statistics on **Table 6.3**).

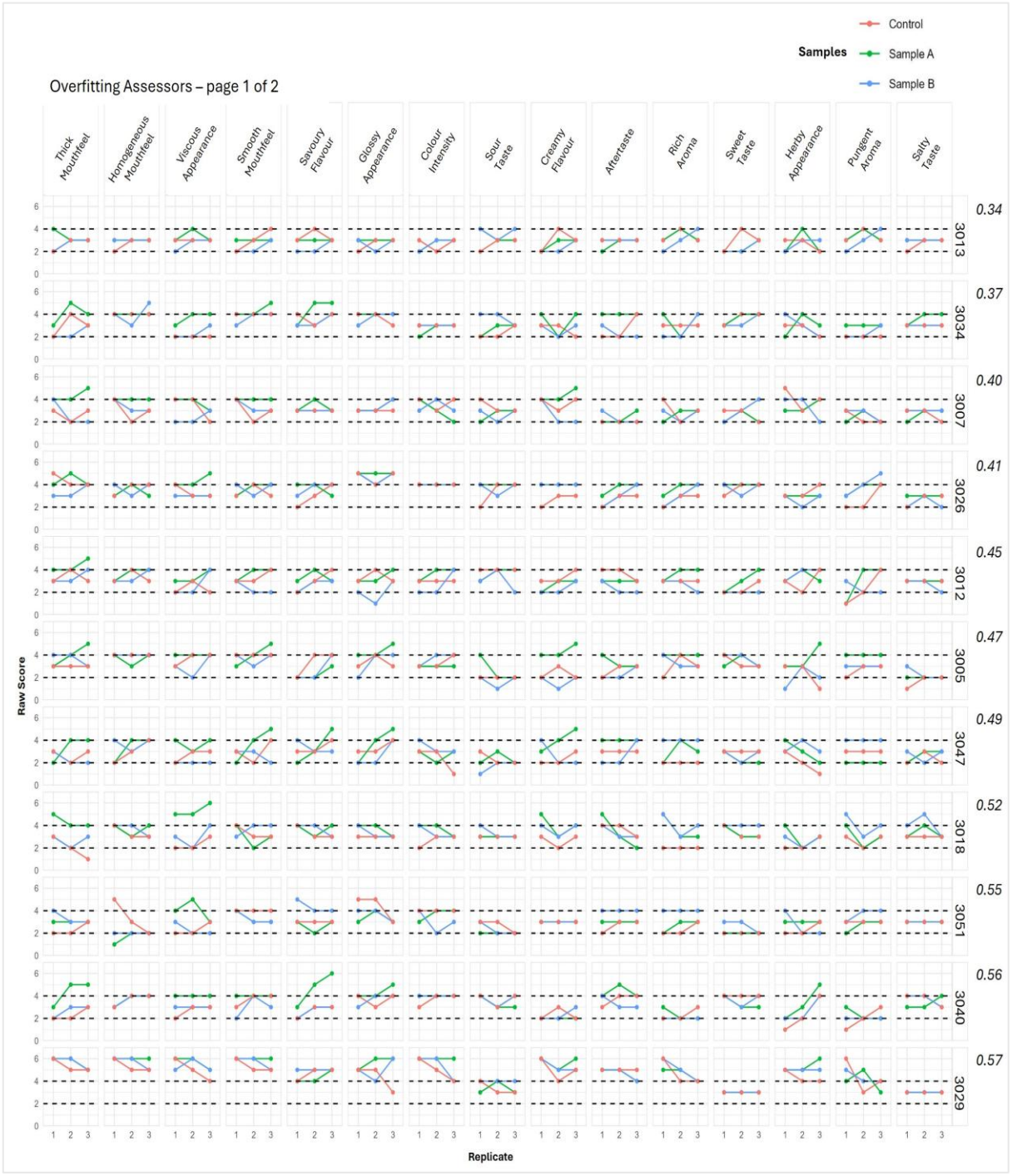


Figure 6.5. page 1. Trellis plots showing the response distribution of raw scores for overfitting assessors. The black dotted horizontal lines representing the central scale categories.

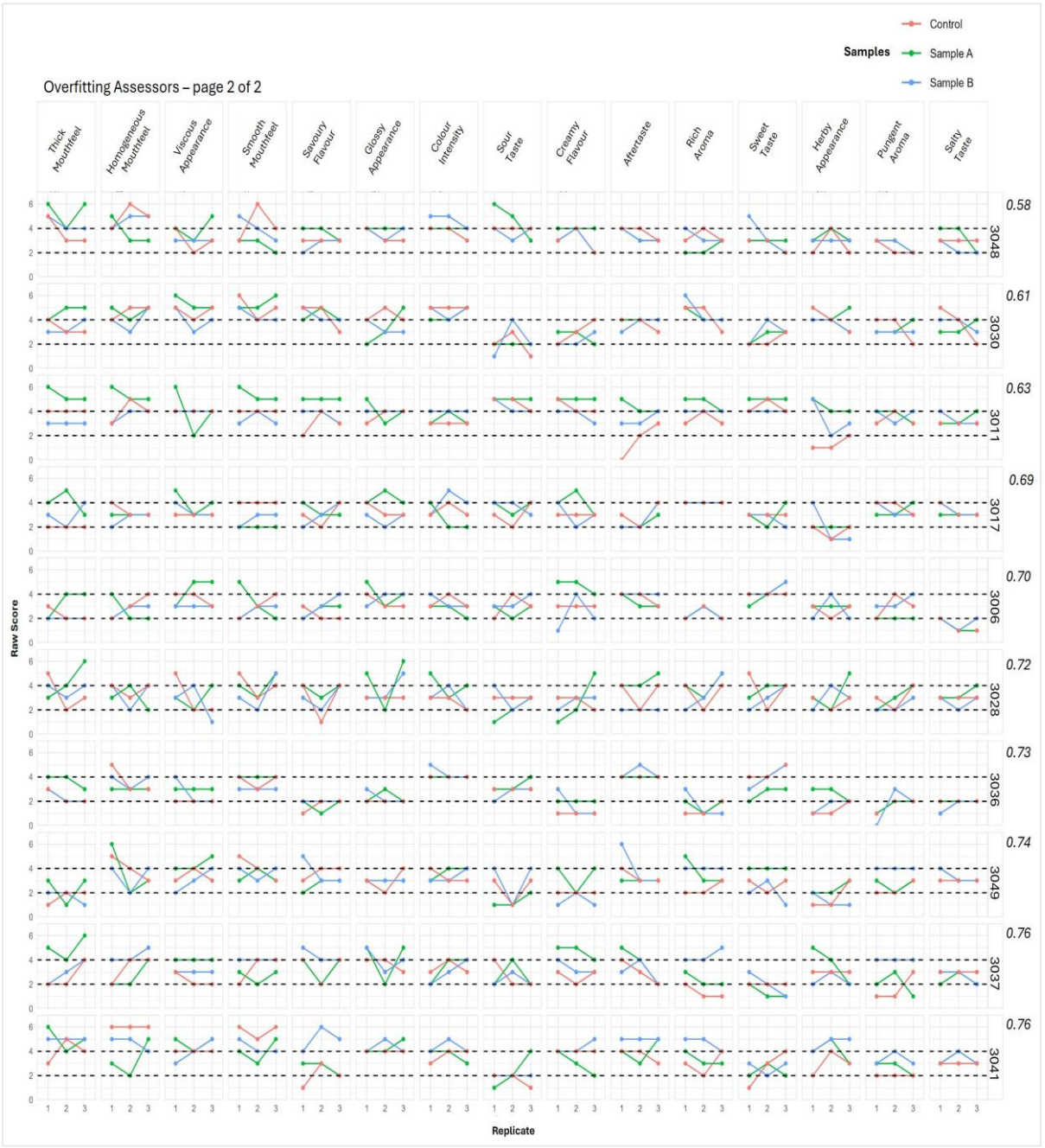


Figure 6.5. page 2. Trellis plots showing the response distribution of raw scores for overfitting assessors. The black dotted horizontal lines representing the central scale categories.

As the figure progresses downward, and OUTFIT Mnsq values increase, the response distributions begin to extend beyond the central margins. Around 0.57, the responses tend to be restricted to other parts of the scale, often away from the centre. This restriction gradually decreases as mean square values approach 0.78. Toward the bottom of page 2 of **Figure 6.5**, where mean square values approach the threshold for overfit, response patterns become less restricted and relatively more consistent.

Although Linacre (2025b) and Myford and Wolfe (2004) note that overfitting assessors are not necessarily poor raters, in a sensory evaluation context, they can substantially affect the validity of the data. Central tendency effects may lead to a lack of discrimination across both samples and attributes, while range restriction may result in poor discrimination across samples alone. In **Figure 6.5,** central tendency is evident where ratings are tightly clustered within the black dotted central margins, while range restriction appears as a narrow band of ratings that mostly fall outside these margins, but still cover only a limited portion of the scale. These rating behaviours are often a consequence of assessors lacking confidence in assigning ratings across the samples, whether due to unfamiliarity with the product range or insufficient training in the use of scale ranges (Lawless & Heymann, 2010; Sipos et al., 2021; Meilgaard et al., 2025). In this study, however, lack of familiarity is unlikely to be the cause, as familiarity with tomato soup was one of the main reasons for selecting these samples for the target participants. In either case, such rating behaviours are problematic, and affected assessors should either receive trainings on confidently using the scale or be removed from the panel. While this study involved an untrained panel, similar issues may also arise among highly trained assessors, potentially indicating a lack of motivation or distraction. This was observed with assessor 2010 in the trained panel discussed in Chapter 5 (section 5.3.4.1: **pg.145**).

Notably, the ratings of assessors 3037 and 3041, both with OUTFIT Mnsq values of 0.76, appear relatively consistent, with fewer crossover interactions. The panel leader may consider retaining such assessors by slightly expanding the cut-off, supported by insights into the type of training required.

As Wright and Masters (1982); Smith (2000), and (Bond et al., 2020) rightly observe, residual fit statistics like OUTFIT Mnsq serve as critical quality control mechanisms. They allow researchers to make informed, interconnected decisions about their data, especially when visual inspection of the full data matrix is impractical. This approach is invaluable for sensory analysts and panel leaders making timely decisions in business contexts.

6.3.6 Key discriminating attributes

The importance of each attribute in distinguishing between the soup samples is presented in **Figure 6.6.** These results are based on data from the 17 selected assessors with relatively consistent rating patterns, compared to the overall panel. The acceptable range for the attributes OUTFIT Mnsq values was calculated* as 0.77 to 1.23, based on 153 total responses per attribute. The findings for the same 17 assessors are discussed in conjunction with the panel interaction plots shown in **Figure 6.7**, the panel ANOVA results (**Table 6.5**), and the assessor responses to the comment section of the DFC questionnaire (**Figure 6.8**), where these assessors indicated which attributes they considered when evaluating the differences between the test samples and the control. Although the DFC comments were collected during the separate occasion for the DFC test, they were provided by the same panel and based on the same set of samples. These comments provide supporting evidence that attributes selected for the TIM AR test were perceivable in the overall product assessment and thus relevant to the assessors. Individual assessor trellis plots are provided in **Figure E 1** for reference.

In **Figure 6.6,** OUTFIT Mnsq values identified *Thick Mouthfeel*, *Homogeneous Mouthfeel* and *Viscous Appearance* as the most discriminating attributes, as they had the highest OUTFIT Mnsq values with positive logit values. This suggests minimal confounding due to assessor confusion or misinterpretation, likely because these attributes were high in intensity and easily perceived. Following these, *Smooth Mouthfeel*, *Savoury Flavour*, and *Herbal Flavour* showed the next highest OUTFIT values. Mouthfeel attributes showed the strongest contributions, consistent with the idea that they are often more readily perceivable than other sensory attributes in this product context. In contrast, attributes such as *Creamy Flavour* and *Sour Taste* had negative logit values and were more difficult to perceive. As discussed in section **4.3.5**, high OUTFIT values for low-intensity attributes often indicate variability, driven more by assessor confusion than by actual product differences. While this pattern suggests that Sample A may have been perceived as the most distinct from the control due to it being thicker and more viscous, this interpretation

* $1 \pm 2\sqrt{\frac{2}{Nr}}$, where Nr (number of responses) for each of the attributes is 153.

is based on perceived differences only, as no rheological measurements were collected to confirm differences in viscosity. This association is supported by the response patterns reflected in the Rasch model's PCAR results discussed in section **6.3.1**, where *Thick Mouthfeel* and *Viscous Appearance* exhibited Local Item Dependency (LID).

These findings were generally consistent with the panel ANOVA (F_{Sample}) in **Table 6.5**, which revealed the same attributes as significantly different ($p<0.05$), and with the assessor responses from the DFC (**Figure 6.8**), except for *Homogenous Mouthfeel*. Contrary to the Rasch findings, *Homogeneous Mouthfeel* was not a significant differentiator across the other metrics. Panel ANOVA revealed only marginal significance across samples ($\alpha = 0.10$), and only a few assessors selected it as a differentiator in the DFC test. The interaction plot (**Figure 6.7**) revealed a pronounced crossover effect, indicating that one assessor rated the samples in a different order than the rest of the panel. Smaller crossover effects were also observed among other assessors. These patterns likely contributed to the elevated OUTFIT Mnsq, by increasing response variability, and may also explain the reduced discriminative power of this attribute (Stone et al., 2012; Raithatha & Rogers, 2018). As discussed earlier, OUTFIT Mnsq represents residual variation, after accounting for individual scale level effects, and reflect variations arising from interaction effects among facets.

Based on the sample composition, it was expected that *Homogeneous Mouthfeel* would differ across samples due to the addition of garlic granules. However, it was uncertain whether this difference would be perceptible and reflected in the tactile perception of homogeneity by the assessors during testing. The definition provided was "Feels the same way throughout" (**Table 6.1**). Assessors may have found this attribute somewhat ambiguous to rate, as several questioned whether higher homogeneity should be scored as higher or lower intensity during the sessions. This underscores the importance of clear, unambiguous descriptions in attribute rating questionnaires (Lawless & Heymann, 2010; Stone et al., 2012; Kemp et al., 2018; Meilgaard et al., 2025).

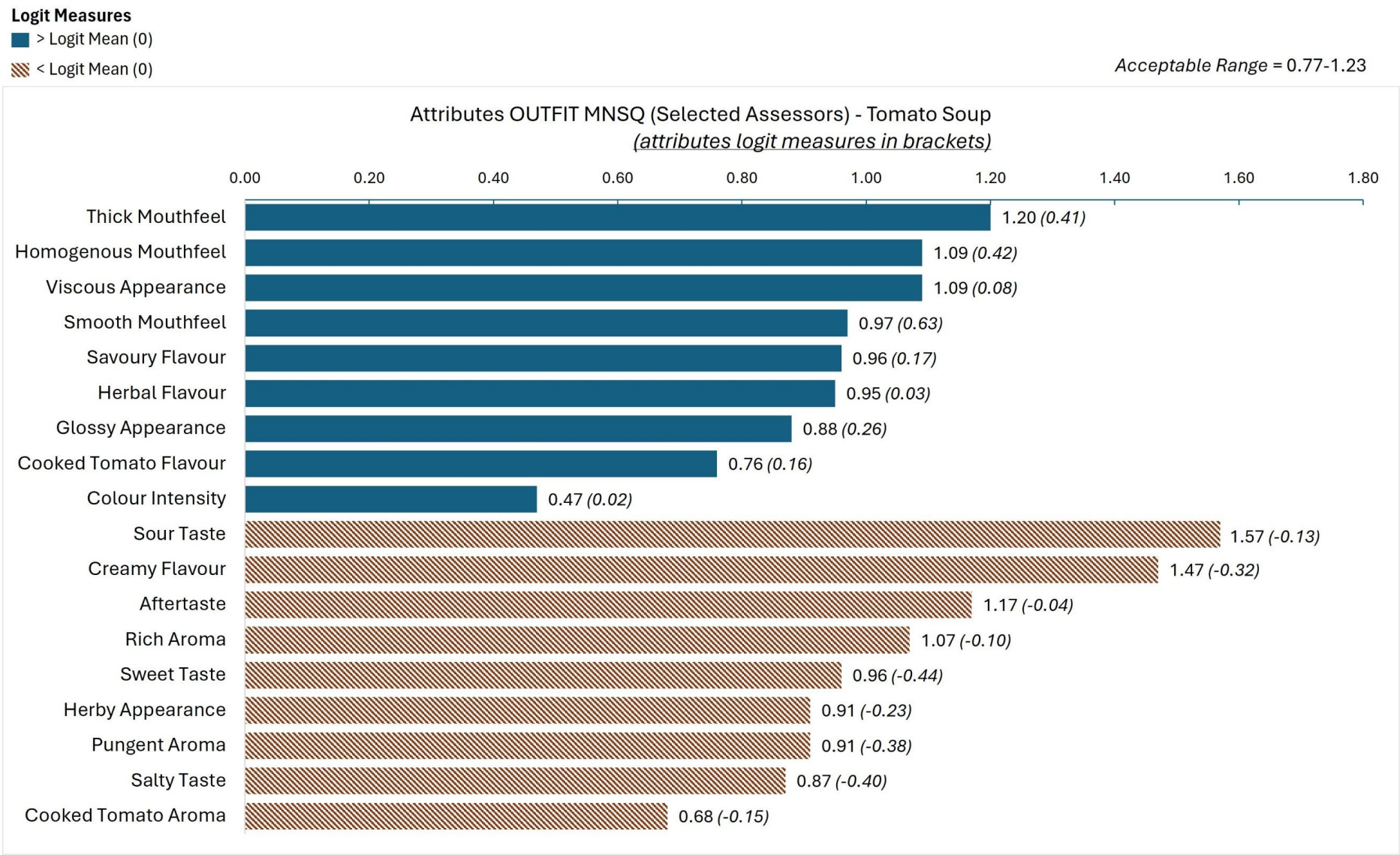


Figure 6.6. Attribute contributions to overall differences between tomato soup samples, based on responses from the selected TIM panel.

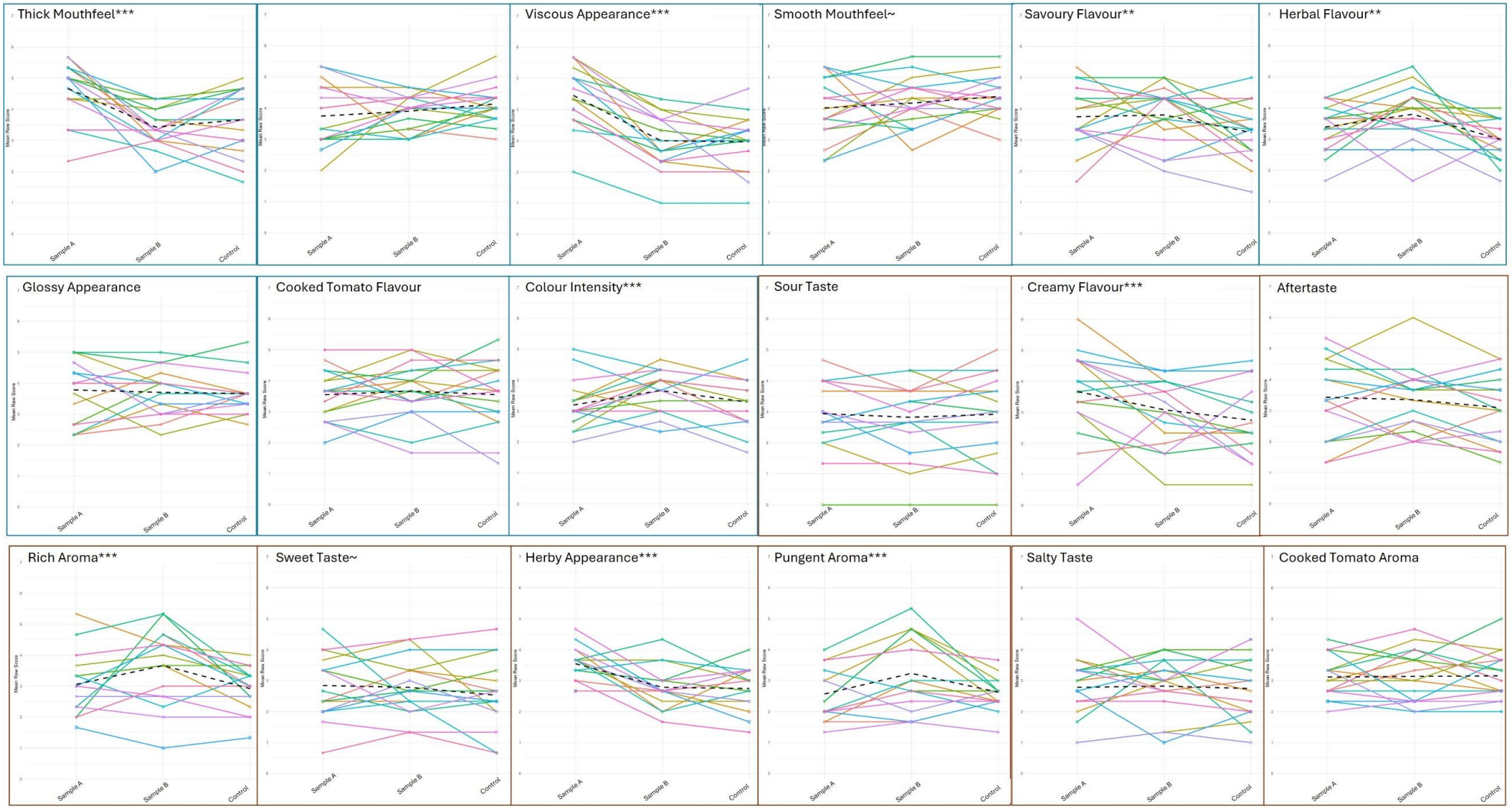


Figure 6.7. Selected TIM panel interaction plots for all attributes.

Plots are arranged in order of decreasing attribute OUTFIT Mnsq values, from top left to bottom right. Attributes are grouped according to whether they fall above or below the logit mean, consistent with their positioning in **Figure 6.6**. Attributes above the logit mean are outlined in blue, and those below are outlined in brown. Attribute titles indicate F-values from panel ANOVA results, with p-value levels significance: <0.001***, <0.01**, <0.05*, 0.10$^{\sim}$; measures with no superscript symbols >0.10.

Table 6.5. Rasch analysis and Raw score ANOVA results for the selected TIM panel on attribute contributions to sample differences.

		Rasch Metrics		Panel ANOVA[1] (N-17)					
	Attributes	**OUTFIT[2]**	**Logit Measure[3]**	**F Sample**	**F Assessor**	**F Assessor x Sample**	**F Repetition**	**F Assessor x Repetition**	**F Sample x Repetition**
+Ve Logit	Thick Mouthfeel	1.20	0.41	**26.83*****	5.19***	**1.80***	1.03	2.12**	0.78
	Homogenous Mouthfeel	1.09	0.42	**2.52**$^{\sim}$	3.03***	**1.77***	1.30	1.72*	0.51
	Viscous Appearance	1.09	0.08	**63.34*****	10.15***	**1.37**	1.65	1.68*	0.17
	Smooth Mouthfeel	0.97	0.63	**3.12**$^{\sim}$	5.92***	**2.43****	1.34	1.81*	0.34
	Savoury Flavour	0.96	0.17	**5.16****	4.90***	**1.68***	0.90	1.12	0.27
	Herbal flavour	0.95	0.03	**7.68****	2.33**	**1.35**	2.46$^{\sim}$	1.50$^{\sim}$	0.63
	Glossy Appearance	0.88	0.26	**0.20**	4.13***	**1.08**	3.73*	1.09	0.46
	Cooked Tomato Flavour	0.76	0.16	**0.34**	8.61***	**1.50**$^{\sim}$	0.80	0.90	1.35
	Colour Intensity	0.47	0.02	**13.38*****	16.03***	**2.8*****	15.55***	2.66***	2.45$^{\sim}$
-Ve Logit	Sour Taste	1.57	-0.32	**0.25**	16.80***	**0.80**	2.55$^{\sim}$	2.06**	0.57
	Creamy Flavour	1.47	-0.13	**12.35*****	9.66***	**2.22****	0.09	1.29	0.31
	Aftertaste	1.17	-0.04	**1.98**	13.17***	**1.24**	1.88	1.70*	2.24$^{\sim}$
	Rich Aroma	1.07	-0.10	**8.25*****	5.65***	**1.64***	0.38	1.43	1.31
	Sweet Taste	0.96	-0.44	**2.69**$^{\sim}$	13.64***	**2.09****	2.43$^{\sim}$	1.91*	0.78
	Herby Appearance	0.91	-0.23	**19.57*****	3.83***	**1.53**$^{\sim}$	7.20**	5.81***	1.24
	Pungent Aroma	0.91	-0.38	**7.78*****	5.97***	**1.30**	2.07	1.37	0.93
	Salty Taste	0.87	-0.40	**0.25**	20.86***	**3.38*****	3.80*	2.90***	2.37$^{\sim}$
	Cooked Tomato Aroma	0.68	-0.15	**0.03**	6.69***	**1.34**	0.10	1.99**	1.63

Attributes are arranged from top to bottom by decreasing OUTFIT Mnsq value and are differentiated based on whether they were located on the positive (+Ve logit > mean) or negative (-Ve logit < mean) side of the logit scale. N signifies the total number of assessors in the panel.

[1] F-values with p-value levels of significance: <0.001***, <0.01**, <0.05*, 0.10$^{\sim}$; measures with no superscript symbols >0.10. n signifies total number of assessors.

[2] OUTFIT Mnsq for attributes indicating whether an attribute's discrimination differs from the average discrimination of other attributes across the samples. Acceptable fit range is 0.77-1.23.

[3] Value of the location of an attribute on the Rasch logit scale: Negative (-Ve) logit values signify low-intensity attributes (below the mean), while positive (-Ve) logit values signify attributes with higher intensities (above the mean).

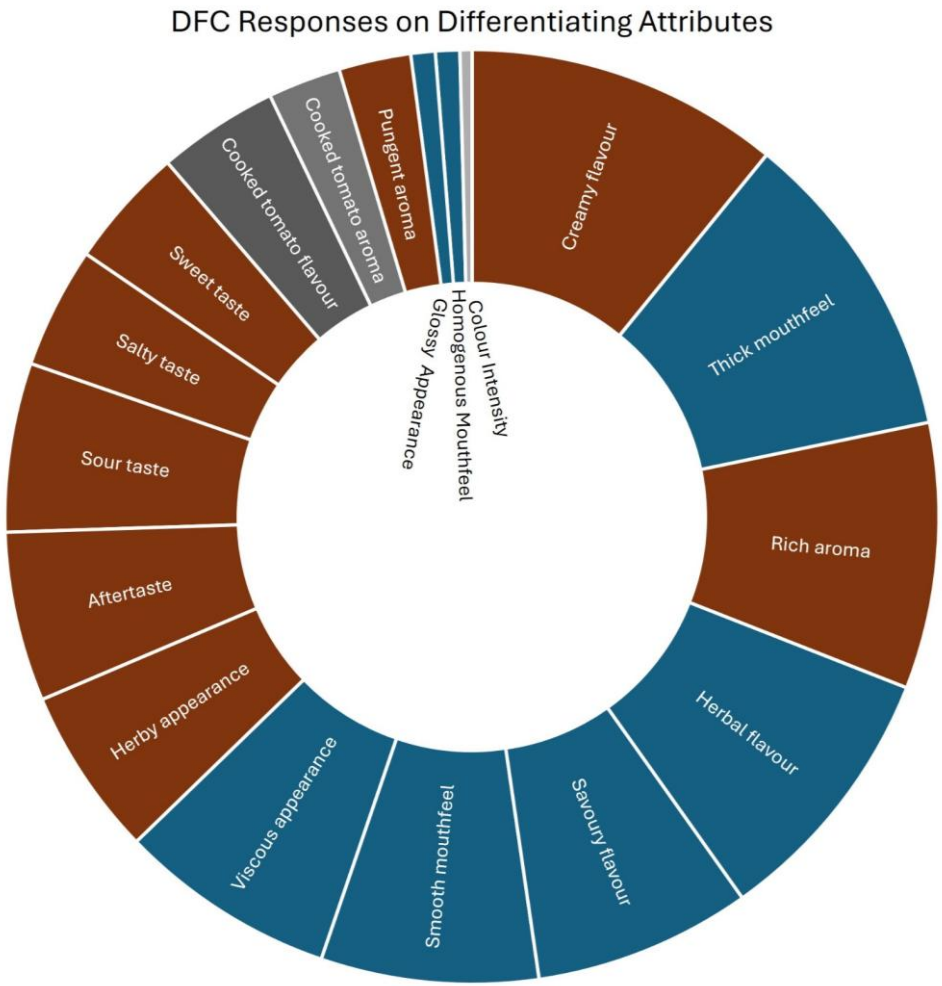


Figure 6.8. Proportion of assessors indicating that a given attribute was considered when evaluating sample differences from the control, based on the questionnaire shown in **Figure C 10.**

Segments are arranged in a clockwise direction, starting from "Creamy Flavour", with proportions decreasing accordingly. Colour shading corresponds to the attribute positions in the OUTFIT Mnsq plot for the TIM method (**Figure 6.6**).**Brown segments** represent low intensity attributes (below the logit mean), while **blue segments** represent high intensity attributes (above the logit mean). **Grey segments** indicate overfit and potentially redundant attributes.

Furthermore, one of the DFC comments cited earlier in section 6.3.4: **pg.196,** mentioned noticeable garlic granules, suggesting that the attribute description "feels the same way throughout" might not have been sufficiently clear. Semantically, even the presence of granules could be interpreted as "feeling the same throughout", leading to potential ambiguity. While this attribute was not found to be a significant differentiator across samples in the ANOVA results and from the DFC comments, Rasch model diagnostics detected irregularities in the response patterns, prompting further investigation. Additionally, this attribute was the only one with a negative point-biserial measure correlation, indicating opposing scale category usage by assessors, which caused inconsistent response patterns and supporting the hypothesis of misinterpretation. Item polarity (discussed in section 3.3.1.4: **pg.58,** and **Table 3.1**) further suggests that the item may be misaligned with the underlying construct and potentially misunderstood by the assessors. This highlights the Rasch model's diagnostic advantage in evaluating product differences and identifying problematic attributes.

Smooth mouthfeel was also revealed as an important contributor both from the DFC responses, and response patterns in the interaction plot, unlike the ANOVA results. The major influence of one assessor, who rated Sample B in the opposite direction from the panel, may have contributed to its high OUTFIT Mnsq value. *Smooth Mouthfeel* and *Homogeneous mouthfeel* were found in the LID analysis (**Table 6.2**) to be statistically related, possibly driven by similar patterns of variation in the responses. Meanwhile, the importance of *Savoury Flavour* and *Herbal Flavour* to product differences were consistent across all the analyses and plots.

Cooked Tomato Flavour, *Glossy Appearance*, *Colour Intensity* and *Cooked Tomato Aroma* were found to be redundant attributes, as their OUTFIT Mnsq values were overfit (below the lower limit of 0.77) except for *Glossy Appearance* which had a value of 0.88. This slightly higher value for *Glossy Appearance* may reflect crossover interactions and poor repeatability, as indicated by the interaction plot. Generally, very few assessors identified these attributes as differentiators in the DFC test, and panel ANOVA showed no significant differences across samples for these attributes.

Interestingly, the only exception among the redundant attributes was *Colour Intensity*, which the ANOVA identified as highly significant ($p < 0.001$), along with

other main effects and most interaction effects. However, based on the sample composition (see photos in **Table B 3**), *Colour Intensity* was not expected to vary meaningfully across the samples, even though slight differences may have been noticeable to some assessors with higher visual sensitivity. This expectation was supported by the overfitting OUTFIT Mnsq value of 0.47, which is well below the lower control limit (0.77), and by the DFC responses, where very few assessors identified it as a differentiating factor. The interaction plots (**Figure 6.7**), and the trellis plots (**Figure E 1**), suggest that the observed differences were mainly due to inconsistent use of the rating scale by the individual assessors. After the Rasch model accounted for these individual differences in severity (i.e. scale level effects), there was likely little true variation in response patterns.

Although the ANOVA model attempts to adjust for differences in scale use through the $F_{Assessor}$ term, the substantial assessor, replicate, and interaction effects, together with the erratic ratings evident in the interaction plots, indicate that these results are not sufficiently reliable to support meaningful conclusions. As noted by Tomic et al. (2007), such adjustments are not always effective when assessors vary in both severity and consistency, and additional methods capable of revealing and accounting for these severity-level effects are required for reliable panel performance monitoring.

Previous research has shown that relying solely on ANOVA results can be misleading when evaluating panel performance and product differences, especially if assessor inconsistency reduces the reliability of the data (Raithatha & Rogers, 2018). To mitigate this, it is recommended that multiple analytical and visualisation methods be combined to ensure reliable interpretations in sensory studies (Tomic et al., 2007; Stone et al., 2012; Ho, 2015; Kemp et al., 2018). These findings therefore highlight the advantage of the Rasch model's diagnostic approach, which provides clearer insight into assessor behaviour and product discrimination and can deliver more confident conclusions with fewer complementary analyses.

The DFC responses revealed that *Creamy Flavour* and *Rich Aroma* were considered by most assessors to be major distinguishing attributes. This finding was corroborated by the Rasch analysis. However, the Rasch model also identified these attributes as relatively lower in intensity and more difficult to perceive,

placing them on the negative side of the logit scale. As discussed in previous chapters, attributes with low intensity scores often reflect confounding effects, such as assessor confusion or inconsistent interpretation, even when the attribute meaningfully contributes to product differences.

When low intensity attributes are also found to be underfit, indicated by OUTFIT Mnsq values above the acceptable upper thresholds, this suggests irregularities in response patterns. This underfit may be due to crossover interactions (where different assessors inconsistently rank the samples) or from extreme rating patterns (where one sample is rated disproportionately lower than others). Despite these effects, the attribute may still be a meaningful differentiator. Typical examples of crossover interactions include Creamy flavour and Rich aroma in this chapter (**Figure 6.7**), where assessors inconsistently ranked the samples. *Orange flavour* ratings from the untrained panel in Chapter 5 (**Figure 5.5** and **Figure 5.6**) is an example of extreme rating patterns, where one sample received disproportionately low scores. Therefore, underfitting low-intensity attributes warrants further investigation, as they may reflect critical sensory characteristics that are harder for assessors to evaluate reliably.

Creamy Flavour showed the second highest underfit with an OUTFIT Mnsq value of 1.57; however, its negative logit value of -0.32 suggested substantial variation in individual interpretations. The interaction plot showed greater agreement among assessors in rating Sample A the highest, but significant crossover interactions between Samples B and the Control, likely contributed to the elevated OUTFIT value. Similarly, the ANOVA revealed significant sample differences ($p<0.001$) and a significant assessor x sample interaction ($p < 0.01$). It was also considered the most differentiating attribute in the DFC test.

The combination of a high underfit score, in a low intensity attribute and pronounced interaction effects, suggests that assessors varied in their interpretation of this attribute, indicating a need for improved training to enhance sensitivity. Difficulties in rating creaminess-related attributes like *Creamy Flavour* have previously been documented. These challenges arise not only from the difficulty in understanding the attribute's meaning (Kilcast & Clegg, 2002) but also from its inherently multisensory nature. Frøst and Janhøj (2007) describe

creaminess as a meta-descriptor encompassing multiple sensory modalities, including taste, texture, mouthfeel, and aroma. More recent work (Corvera-Paredes et al., 2022) confirms that creaminess perception in foods depends on complex interactions among viscosity, lubrication (tribology), food matrix structure, and salivary processes. Although assessors were instructed to evaluate creamy flavour specifically, the multidimensional combination of these sensory characteristics may have led different assessors to focus on different aspects, such as mouthfeel thickness or richness of flavour.

Similar challenges have been noted for other mouthfeel-related attributes, with research showing that attributes such as astringency consist of several perceptual sub-qualities, making them more difficult to evaluate consistently (Wang et al., 2020). Furthermore, individual physiological differences contribute additional variability, as variation in salivary composition has been shown to affect the perception of astringency sub-qualities and influence how individuals experience and rate these mouthfeel sensations (Wang et al., 2021). Oral processing and salivary dynamics have been shown to shape texture and mouthfeel perception (Stokes et al., 2013), with lubrication properties playing a particularly important role (Boehm et al., 2020). These findings reinforce that individual biological differences can contribute to inconsistent ratings for complex, multisensory attributes such as creaminess. This multisensory complexity likely explains the variability and inconsistencies observed in the ratings for *Creamy Flavour*.

Rich Aroma was also identified as a challenging attribute for assessors to rate. Its OUTFIT Mnsq value of 1.07, approaching the upper acceptable threshold of 1.23, along with a negative logit value of -0.10, suggested that the attribute was perceived at a relatively low intensity and that assessor responses varied considerably. This was further supported by the interaction plot, which showed significant differences across samples along with pronounced crossover interactions involving all three samples.

One possible explanation for this inconsistency is the cultural variation among assessors, which may have influenced their interpretation of the attribute definition. *Rich Aroma* was defined as “Combination of multiple ingredients creating a deep and full aroma. For example, well-seasoned food”. The panel

included assessors from Western Europe, Mediterranean Europe, South America, and South Asia, regions with diverse culinary traditions and sensory expectations. Cross-cultural research has consistently shown that cultural background shapes how individuals perceive, evaluate and describe aromas. Recent studies have shown that cultural background plays a substantial role in how aromas are perceived, categorised and described. For example, Sharma (2023) demonstrated that odour vocabulary and conceptualisation vary significantly across cultures, influencing how individuals interpret aromatic cues. Majid et al. (2018) similarly reported pronounced cross-cultural differences in olfactory naming, discrimination and perceptual organisation across diverse linguistic groups. Earlier work by Pangborn et al. (1988) found that regional aroma-liking varied significantly across geographic areas, likely because of differences in traditional food habits and aroma availability rather than purely sensory threshold differences. Together, these findings support the likelihood that cultural variation contributed to the inconsistent ratings of *Rich Aroma* in this study.

For some assessors, Sample A was perceived as having the richest aroma, likely also influenced by the enhanced complexity in aroma and flavour perception, caused by the increased viscosity when double cream was added (as discussed in section 6.3.1: **pg.194**). For others, Sample B was rated the highest, which aligns with the sample design, since the addition of garlic was expected to increase both the *Savoury Flavour* and *Rich Aroma* attributes. **Table E 2** also confirms this, showing that Sample B had the highest average intensity score for *Rich Aroma* and was significantly different from the other samples ($p<0.05$).

As discussed earlier in section **6.3.1**, LID analysis revealed a statistical relationship between *Rich Aroma* and *Savoury Flavour*, supporting the connection between these two sensory attributes. However, according to **Table E 2**, Sample A had the highest intensity for *Savoury Flavour*, which was defined in the questionnaire as "Rich, spicy flavour associated with vegetable or meat broth". Despite this relationship, results indicated that *Savoury Flavour* was easier for assessors to rate consistently than *Rich Aroma*. This suggests that *Rich Aroma* was not only more culturally variable, but also perceptually more complex, possibly due to it being perceived at a relatively lower intensity. The finding that *Savoury Flavour* was rated

highest in Sample A further highlights the alterations to other sensory attributes, and sensory interactions resulting from the addition of extra cream.

Several studies have continued to emphasise the importance of considering cultural influences when designing and interpreting results from sensory and consumer research (Muñoz, 2002; Harrington, 2005; van Zyl & Meiselman, 2016; Ares, 2018; Hort, 2024).

Notably, except for *Herby Appearance* and *Creamy Flavour*, the low intensity attributes consisted entirely of taste and aroma characteristics, including aftertaste. It is well established in the sensory literature that attributes related to smell and taste, particularly those perceived at low intensities or near detection thresholds, are often difficult for assessors to rate reliably. This underscores the training requirements to improve sensitivity to these attributes within a given product range (Lawless & Heymann, 2010; Kemp et al., 2018; DLG, 2020; Meilgaard et al., 2025). The interaction plots supported the Rasch analysis findings, revealing multiple magnitude and crossover interactions in the rating patterns. However, the OUTFIT Mnsq values increased with more crossover interactions because the Rasch model had already removed the effects of individual differences in scale use, which are reflected by the magnitude interactions (Stone et al., 2012).

Herby Appearance was another challenging attribute to assess, but was suspected to have been influenced more by the sample design. The soup base for Sample A was a cream of tomato soup without basil, unlike the other samples. Dried chopped basil leaves were added while reheating the sample, shortly before evaluation. As a result, some assessors may have received samples with more visible basil specks than others, introducing inconsistencies in the perception of this visual attribute.

In summary, findings from earlier chapters have demonstrated how the Rasch-based TIM approach offers a streamlined diagnostic framework to identify key attributes driving product differences and to highlight individual assessor behaviour. By revealing unexpected and inconsistent response patterns, the model enables deeper examination of attribute perception and potential interactions, informed directly by observed panel ratings, without requiring multiple separate statistical techniques as in traditional analyses.

However, it is important to clarify that MFRM adjusts only for systematic differences in assessor severity (consistent tendencies to assign higher or lower ratings) and does not correct for broader individual differences in scale interpretation, attribute understanding, or physiological variation. Thus, the model improves comparability but does not eliminate the need for careful attribute definitions and panel training. Unexpected variations, whether systematic or inconsistent, are flagged through diagnostic outputs such as PCAR and residual fit statistics (e.g., OUTFIT Mnsq), which support informed interpretation rather than definitive explanations.

While DFC achieved higher separation and reliability through its simpler comparative design, TIM offers distinct advantages when research goals extend beyond quantifying overall difference. It simultaneously identifies which specific attributes drive product differences, their contribution to the overall difference, and sources of measurement variability such as assessor inconsistency or confusion. The choice depends on whether only overall difference quantification is needed (DFC) or comprehensive diagnostic insights into both products and panel performance are required (TIM). These interpretations are most reliable when there is acceptable fit to the model, with visualisations of raw data helping to clarify patterns and support conclusions.

6.4 Limitations of the study

This study aimed to explore the transferability of the Rasch-based approach within a simulated context, designed to reflect settings commonly found in sensory quality programmes. A panel of untrained assessors was employed to investigate the potential of the Rasch model to identify individuals capable of performing at a trained assessor level.

Panel performance constraints

While some limitations identified in the previous two studies (particularly regarding sample and attribute choices discussed in Chapters 4 and 5) were successfully addressed, the use of an untrained panel continued to pose challenges due to inherently inconsistent ratings within each assessor (as shown in **Figure E 1**). As demonstrated by the trained panel in Chapter 5**:** section **5.3.4.1**, Rasch-based diagnostic insights are considerably more informative when the panel operates at a

relatively standardised level. This is because the model evaluates rating behaviours and flags unexpected variations relative to the overall panel performance (Myford & Wolfe, 2004).

The subset of better-performing assessors (n=17) selected from the full untrained panel (n=54) exhibited better consistency than the group overall, but still displayed some inconsistency, likely attributable to limited sensitivity to product attributes. Consequently, trends for identifying problematic assessors, such as through SR/ROR statistics, were not as clear-cut as was shown with the trained panel in Chapter 5.

Lack of instrumental analysis to verify product characteristics

Additionally, as with the previous studies, no instrumental analysis was conducted to verify the actual presence or intensity of the attributes evaluated. For instance, rheological measurements for viscosity, and particle size analysis for smoothness, could have provided objective confirmation of the perceived sensory differences. All findings are therefore based on panel ratings, and conclusions regarding the physical product characteristics underlying these differences remain interpretations rather than instrumentally verified properties.

A trained panel sensitive to the attributes used in this study would likely have more clearly demonstrated the strength and impact of Rasch-based diagnostics for monitoring assessor performance. Nevertheless, the study yielded valuable insights into rating behaviours and attribute contributions, as reflected in OUTFIT Mnsq values. It offered nuanced diagnostics of both systematic variations (through PCAR) and non-systematic variations, and yielded more accurate assessments of product differences by accounting for one of the most significant sources of individual variation: the idiosyncratic use of rating scales (Linacre, 1994; Lawless & Heymann, 2010; Meilgaard et al., 2025). These findings underscore the potential of the MFRM, even when applied in contexts involving untrained assessors.

6.5 Significance of study

This final study extends and consolidates the findings from previous chapters by examining the adaptability and diagnostic capabilities of the MFRM and TIM approach in sensory difference testing across varied contexts. The results demonstrate that the MFRM offers diagnostic efficiency, particularly in situations where manual inspections

of data matrices or multiple independent statistical tests would be time intensive. By modelling products, assessors, attributes, and scale categories simultaneously, the model enables unexpected or inconsistent response patterns to be detected in a single analytical step. This offers practical benefits for sensory analysts, panel leaders, product developers, and quality managers in organisations with established sensory programmes and frequent assessment needs. The ability to identify which attributes drive perceived differences (as rated by the panel), and which contribute noise, can support more targeted reformulation, optimisation, and shelf-life decisions, while helping distinguish genuine product changes from panel drift or rater instability.

While Chapters 4 and 5 demonstrated these capabilities across both trained and untrained panel contexts, this chapter specifically investigated whether MFRM diagnostics could support assessor selection and recruitment decisions in sensory quality programmes. Building on the finding from Chapter 5 that fit statistics are relative to panel performance, this study examined whether identifying assessors with more stable response patterns within an untrained group (i.e., acceptable fit within the panel) could improve panel discrimination and reliability. The results showed that diagnostic clarity scales with panel quality: highly informative with trained assessors (Chapter 5), moderately useful for identifying relatively stable performers among untrained groups for potential recruitment (Chapter 6), but with important caveats regarding fit statistics interpretation in poor-quality panels (discussed in section Chapter 5**: pg.164**). This progression clarifies both the utility of MFRM for practical assessor screening and the boundary conditions for its application in sensory quality control settings.

Compared with traditional ANOVA-based approaches, the Rasch framework offers several strengths while remaining complementary rather than substitutive. ANOVA remains highly suitable for routine product comparisons with well-established, well-trained panels. However, the MFRM provides a unified analytical process that simultaneously accounts for systematic rater severity and offers integrated diagnostics not accessible through ANOVA. These include finer-grained insight into which attributes discriminate products effectively, which assessors deviate from panel expectations, and how rating scales are being applied.

Tomic et al. (2007) emphasised that robust methods are needed for compensating for panel drift, correcting for individual assessor level and range differences in scale use, and computing weighted sample estimates. This is particularly relevant given the substantial resources required to train and maintain calibrated descriptive panels. Panel training commonly ranges from 10 to over 100 hours (Djekic et al., 2021), and in some cases extends over several months (Lestringant et al., 2019), with ongoing calibration and drift monitoring representing continuous operational demands. The MFRM contributes to addressing these needs by adjusting for individual assessor severity differences (scale level correction), flagging assessors with problematic scale use, and pinpointing scale range effects like central tendency and range restriction through fit statistics, enabling targeted intervention for range correction. When time-related facets like days, months or sessions are included in the model design, it can also reveal changes in rating patterns over time that may indicate performance drift. These capabilities offer practical operational benefits for panel management. Because the model adjusts for consistent differences in severity, panel leaders do not need to devote training time to enforcing identical rating standards across assessors, allowing them to focus instead on attribute understanding and meaningful interpretation of intensity levels.

Additionally, integrated diagnostics such as fit statistics and SR/ROR correlations allow rapid identification of specific assessors who require targeted intervention, facilitating timely correction of issues such as scale misuse, inconsistency or attribute confusion. While the magnitude of time savings will vary by organisation, these diagnostic capabilities indicate meaningful opportunities to streamline panel maintenance and reduce the cumulative burden of ongoing calibration.

While the MFRM offers clear advantages, its implementation also requires consideration of practical constraints. The method involves specialised software and methodological training, which may limit adoption among stakeholders unfamiliar with probabilistic models. Some resistance is likely in organisations that rely on established workflows centred on ANOVA. However, once implemented, the MFRM can reduce downstream workload by consolidating multiple analyses into a single model and by minimising the need for extended panel retraining. As sensory analysis software continues to evolve, the integration of Rasch-based tools into

commercial platforms may further lower barriers to adoption by automating model fitting and diagnostic reporting. The balance between initial setup effort and long-term diagnostic gains favours adoption in organisations with sustained quality monitoring needs and sufficient technical capacity to support initial implementation.

An additional caveat concerns attribute selection and the assumption of unidimensionality. As shown in Chapter 5: **pg.159** , when one product among those being compared exhibits an opposing attribute profile (e.g., Brand B with the lowest Orange flavour but highest Milky flavour ratings), the latent *Overall Difference* estimate becomes distorted through cancellation effects on the logit scale, reducing product separation clarity. This occurs because opposing attributes do not align along a single underlying construct, violating the unidimensionality assumption. Careful attribute selection is therefore critical to ensure attributes collectively represent a coherent sensory dimension. This consideration is particularly important for untrained panels, where inconsistent scale use can amplify distortions from poor attribute selection.

Across the three studies, these diagnostic capabilities were delivered through a consistent analytical framework summarised in **Figure 6.9**. Key tools including the Many-Facet Wright map, separation statistics, rating scale category statistics, Principal Component Analysis of Residuals (PCAR) and Local Item Dependency (LID) analyses, residual fit statistics (OUTFIT mean squares), and Single Rater Rest-of-Rater (SR/ROR) correlations collectively provide a coherent view of product discrimination, assessor behaviour, attribute functioning, and potential sensory interactions. Rather than requiring multiple separate analyses, these tools operate within a single model and offer a rapid, integrated interpretation of the underlying data structure, one of the key practical advantages of the MFRM over traditional approaches.

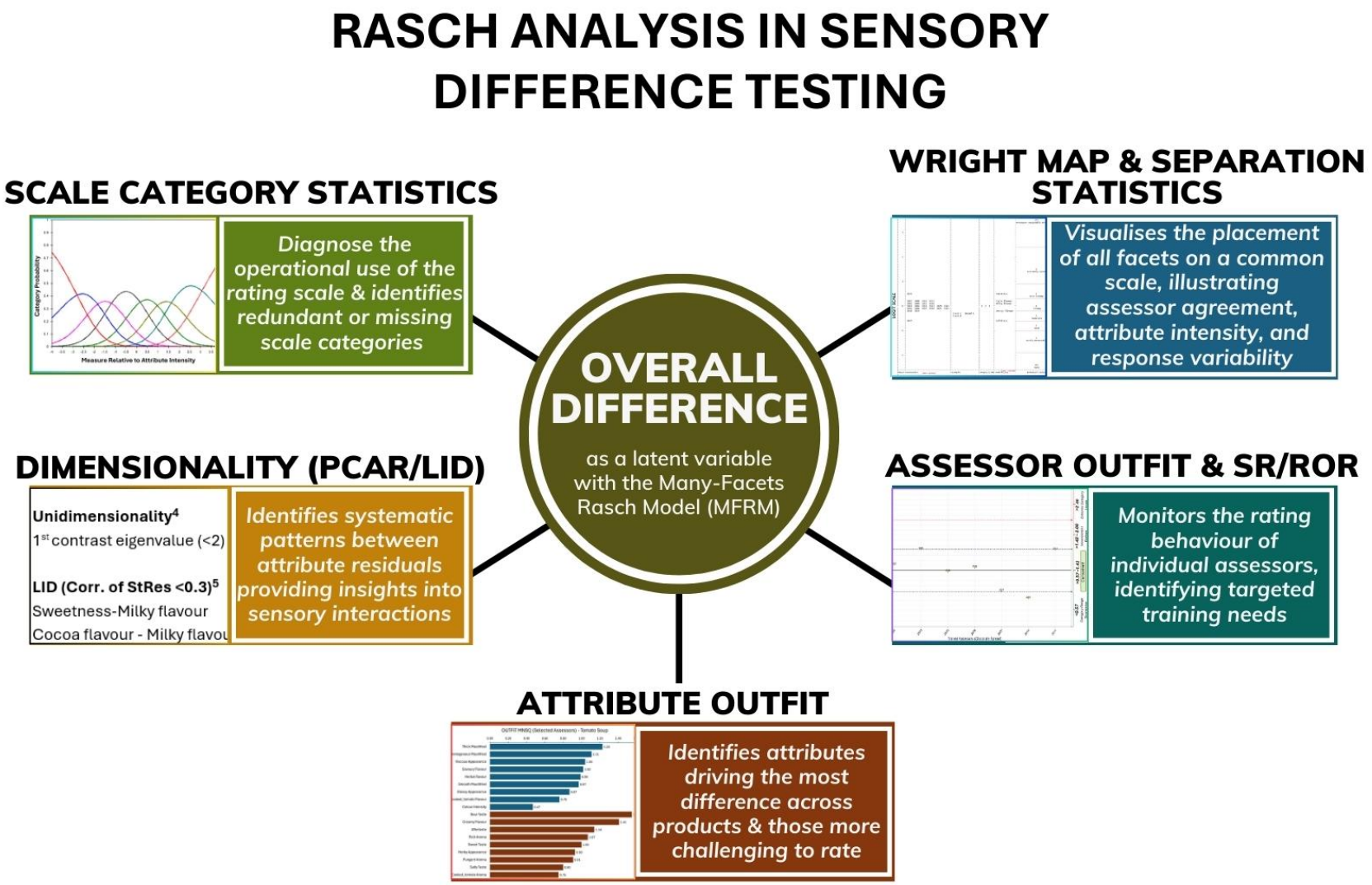


Figure 6.9. Schematic summary of the application of Rasch analysis in sensory difference testing.

The diagram outlines the diagnostic tools employed across the three study contexts, highlights the specific insights each tool provided, and demonstrates how they collectively support the use of the Many-Facet Rasch Model (MFRM) as a robust analytical framework for measuring overall difference between products.

Chapter 7

Overall Discussion

7.1 Summary of findings

This research explored the application of the Many-Facet Rasch Model (MFRM) in sensory difference testing with the goals of estimating an overall difference latent variable from multiple sensory attributes, evaluating the model's ability to account for individual differences in rating scale usage, and exploring the diagnostic potential of Rasch outputs to improve sensory data interpretation and panel performance.

The findings across the three study contexts demonstrate that the model effectively addressed the research aims. Specifically, the MFRM successfully estimated a single latent measure of *Overall Difference*, by combining intensity ratings across attributes, and this measure was comparable to the holistic overall difference score derived from the DFC test. The inclusion of assessors as a distinct facet in the model, allowed the analysis to account for individual differences in rating scale usage, by estimating each assessor's severity or leniency. This contrasts with the conventional ANOVA, which relies on aggregated data and does not model individual-level variability. Additionally, Rasch-based diagnostics such as fit statistics, Wright maps, and PCAR, uncovered inconsistencies and the underlying structure of the response data.

As a result, the Rasch-based approach provided valuable, more detailed insights than traditional methods, all within a single, integrated analysis. This offered enhanced practical value in several ways: (1) identifying inconsistent assessors early for targeted retraining and (2) clarifying which attributes contributed meaningfully to sample discrimination, both using the fit statistics; and (3) enabling rapid and intuitive interpretation of results through Wright maps, which visually place assessors and attributes on the same measurement scale, making outputs easier to understand for non-technical stakeholders. These advantages support more efficient panel management, clearer insights for product development, and better-informed quality control decisions across the broader business context. While

implementation requires specialised software and training in Rasch principles, once established the unified analytical framework can reduce the need for multiple separate analyses and can streamline panel management. Moreover, integration of Rasch capabilities into commercial sensory software platforms could further reduce implementation barriers and increase accessibility through automated model fitting and diagnostic reporting.

7.1.1 Estimating overall difference: Integrating quantitative & qualitative insights

The Rasch approach yielded a single measure - Total Intensity Measure (TIM) for each sample using observable ratings across several selected sensory attributes. This measure represents the latent variable of *Overall Difference*. While post hoc tests produced conclusions similar to the traditional DFC overall difference test, the Rasch method offered enhanced clarity by not only helping to focus assessor attention on critical attributes of interest, but also generating valuable qualitative, diagnostic insights such as:

- Revealing the relative perceptibility of different attributes using the Wright map, indicating which attributes were most easily and prominently perceived, and which were the least noticeable to assessors across all samples.
- Identifying the key drivers of the perceived product differences using OUTFIT Mnsq statistics, highlighting attributes that contributed most reliably to sample discrimination, and
- Clarifying whether observed differences stemmed from true sample variation or from panel inconsistency and assessor confusion, also through the OUTFIT Mnsq statistics for attributes.

Previous studies have attempted to enrich the DFC method either by targeting specific attributes (Higgins & Hayes, 2020; Cela et al., 2023), or by incorporating open comment sections or CATA questions for qualitative insights (Rogers, 2017; Compusense, 2020). However, the proposed Rasch approach delivers more comprehensive, integrated quantitative and qualitative insights in a single, streamlined analysis, eliminating the need for multiple, separate tests.

Furthermore, ANOVA-based methods operate at an aggregate level, and do not account for individual rating behaviours (Næs, 1990; Romano et al., 2008; Næs et

al., 2010; Hannum et al., 2019). While ANOVA can detect differences between samples and assessors' overall discriminatory ability, it cannot pinpoint which assessors struggled with specific attributes or identify sources of disagreement at the individual level. In contrast, Rasch modelling, accounts for individual rating styles by adjusting for severity or leniency biases and provides attribute-level fit diagnostics (e.g., OUTFIT Mnsq) that help distinguish true product differences from those driven by assessor inconsistency or confusion. This enhanced interpretability supports:

- More targeted assessor training based on areas of inconsistency or confusion.
- Elimination of redundant or non-discriminative attributes, and
- A sharper focus on key drivers of product differences during analysis and reporting.

Finally, response dependency analyses revealed patterns of co-variation among attributes, which may suggest potential relationships such as ingredient interactions or attribute synergies perceived by the assessors. While these correlations do not establish causation, they offer valuable starting points for further investigation and can inform product development efforts, providing useful cues for product formulation, optimisation, and quality control.

7.1.2 Comparing rating behaviours of trained and untrained assessors

Regarding individual variability (outlined in **Table 2.1**), results from Chapter 5 confirmed that trained panels are generally more sensitive to product attributes and consistent in their ratings. The trained panel in this study consisted of expert assessors with extensive sensory profiling experience for a global chocolate manufacturer. However, even the experts occasionally exhibited inconsistencies. Their tendency to give more conservative ratings, as observed with the *Orange flavour* attribute, is hypothesised to reflect the cumulative influence of feedback received over multiple prior training sessions (Castura et al., 2005; Raithatha & Rogers, 2018). Notably, the panel had not received specific training on chocolate-orange spreads prior to this study, which may have contributed to their inaccurate ratings for *Orange flavour* in the chocolate spread samples. This aligns with Chollet

et al. (2005) and Ares and Varela (2017), who argue that the perceptual acuity of trained assessors does not necessarily generalise to stimuli outside their training.

In examining scale-use bias, a well-documented challenge in sensory profiling (Næs, 1990; Romano et al., 2008), the MFRM in this study adjusts for idiosyncratic scale usage (level effects), but still captures individual deviations through residual fit statistics. Assessor OUTFIT Mnsq identified various scale-use biases: overfitting assessors showed range restriction or central tendency, while underfitting assessors exhibited erratic behaviour or extreme responses. Untrained panels predominantly displayed high variability, reflecting differing experience levels and lower sensitivity to some attributes. Chapter 6 highlighted instances, where the lack of a shared frame of reference for scaling among untrained assessors, led to inconsistent interpretations of complex attribute descriptors like *Creaminess* and *Homogeneous*, similar to inconsistencies in attribute understanding and scale use reported by Antmann et al. (2011) and Ares et al. (2011).

As observed by Worch et al. (2010) and Xiangli et al. (2024), the trained and untrained panel in Chapter 5 generally produced similar directional results (as shown on the Wright maps). However, while there was a significant difference between at least one of the samples for the trained panel, this was not the case for the untrained panel, as their internal inconsistencies (i.e. the variability effect from crossover interactions within themselves), resulted in the loss of discrimination (Stone et al., 2012). The Rasch model corrected for this internal variability, filtering it out as unsystematic noise rather than meaningful differences. This demonstrates the model's strength in objectively managing variability, albeit conservatively, such that only sufficiently systematic differences, beyond severity or leniency in scale use, are deemed reliable. Raithatha and Rogers (2018) noted that panel results must be consistent enough for the panel mean to represent genuine product differences.

In business contexts, the Rasch models' objectivity is advantageous. Another instance from Chapter 6 was how traditional ANOVA on raw scores indicated a significant difference in *Colour intensity*, including assessor and interaction effects. In contrast, the Rasch model, after adjusting for scale-use effects, found no meaningful product differences, aligning better with DFC results. This instance

highlights how the conventional approach to analysing aggregated sensory data can be inefficient.

That said, in agreement with Ares and Varela (2017), Meiselman (2013), and Barton et al. (2020), the choice between trained and untrained assessors should be context-dependent. Trained panels are not infallible, and untrained assessors, while lacking consistency, bring valuable familiarity with the product of interest. When combined with Rasch-based adjustments for individual rating style, untrained panels can yield consumer-relevant insights and effectively substitute for trained panels, offering significant resource savings, by reducing the need for extensive training in some industrial settings. Therefore, untrained assessors can represent a practical and valid option in many cases, especially when resource constraints or the nature of the product evaluation make trained panels less feasible.

7.1.3 Diagnostic insights into the use of rating scales

This study employed a category-labelled (ordinal) scale, whose anchors were adapted from the Labelled Magnitude Scale (LMS). The rationale was to provide well-established intensity descriptors while maintaining the simplicity of categorical rating, and to leverage the Rasch model's ability to convert ordinal scores to interval measures. The LMS anchors were chosen to enhance discriminative capacity and prevent ceiling effects. However, no specific training on the use of the scale in this study was provided for either of the panels. Additionally, this study did not empirically compare the adapted LMS format against traditional category labels or unstructured line scales to determine whether it actually improved usability or discrimination.

Wright maps generated across all studies visualised individual scale categories as threshold ranges, indicating transition points between rating categories. Slightly unequal distances between adjacent categories, widening toward the extremes, are characteristic of category-ratio scales like the LMS (Green et al., 1993), the Borg Scale (Borg, 1982), and the generalised Labelled Magnitude Scale (gLMS) (Bartoshuk et al., 2005). Notably, the end category “Strongest imaginable oral sensation” was the least used across all studies. Prior research by Schifferstein

(2012) and Hayes et al. (2013) has highlighted that untrained assessors often struggle with abstract anchors like “strongest imaginable”, resulting in scale compression and underuse of the upper categories. Interestingly, the trained panel also underutilised the “extremely strong” and “strongest imaginable” upper categories while consistently using more middle categories. Given the lack of scale use training, this pattern likely reflects the actual intensity range of the product attributes, or the trained panel’s unfamiliarity with the LMS scale, as they were more accustomed to unstructured line scales. This demonstrates that sensory training does not necessarily generalise across different scale formats, highlighting the need for scale-specific familiarisation even with experienced panels.

The scale category diagnostic information produced by the MFRM proved valuable for identifying these usage patterns. The model successfully identified redundant categories (such as the "Strongest Imaginable" category across all three studies) and revealed patterns in scale use that inform both scale refinement and training needs. However, whether the adapted LMS format actually enhanced discrimination compared to traditional labels or reduced rating variability compared to line scales cannot be determined from this study, as comparative data were not collected.

For manufacturers, the Rasch approach can support the design of long-term sensory quality programs, by helping to optimise rating scales. In this study, the MFRM allowed ordinal category scales to be converted to interval-level measures while providing diagnostic insights into how assessors actually applied the scale. Rasch analysis adjusts for individual differences in how assessors interpret and apply scale categories (in terms of severity and leniency), reducing, but not eliminating, the need for intensive training to ensure consistent attribute understanding and rating behaviour, which remains a persistent challenge in sensory and consumer research (Kemp et al., 2018).

Ultimately, the Rasch-based framework aligns with Meiselman's (2013) call to prioritise efficiency and usability of sensory rating scales, over ongoing debates about whether a scale is inherently “good” or “bad”. However, this study demonstrates that even with Rasch calibration, scale design choices, particularly anchor terminology and the provision of scale-specific training, matter for practical application, and warrant systematic empirical investigation in future research.

7.2 Limitations

The studies presented in Chapters 4 to 6 demonstrated the diagnostic capabilities of the MFRM across different panel types and product contexts. Several limitations warrant consideration when interpreting findings and planning future applications.

Absence of instrumental verification

Across all studies, no instrumental analysis was conducted to verify sensory attributes. Tests such as headspace GC-MS could have confirmed volatile compound profiles underlying perceived aroma and flavour attributes (e.g., orange, herbal, cooked tomato), HPLC or titration methods could have quantified non-volatile compounds contributing to taste perceptions such as sweetness (sugars), sourness (acids), and saltiness (sodium content), texture analysis and rheometry could have verified structural properties such as viscosity in chocolate spreads and tomato soups or sponginess in Jaffa cakes, and particle size analysis could have confirmed perceptual differences in smoothness of the soups. This limits the ability to distinguish genuine product variation from perceptual variation or measurement error. While the MFRM provides robust measurement of perceptual differences, conclusions about the physical product characteristics underlying those differences remain interpretations rather than instrumentally verified properties.

Model assumptions and validity

The validity of Rasch model-derived insights depends on key assumptions being met. The model estimates a single latent "Overall Difference" variable, which assumes all attributes contribute to a unified sensory dimension (unidimensionality). When products exhibit opposing attribute intensities, as observed with Brand B in Chapter 5 (the only sample with no added orange flavouring, resulting in low *Orange flavour* but high *Milky flavour* ratings compared to other samples), this assumption may be violated, causing cancellation effects on the logit scale that mask true product differences. Future applications should carefully select attributes that align along a coherent sensory dimension. When one product among those being compared has a distinctly different attribute profile, preliminary examination of raw data and attribute patterns can inform whether a unidimensional overall difference measure is appropriate, particularly when using untrained or inconsistent panels.

Additionally, fit statistics are benchmarked against the response patterns within a panel, not external standards. Acceptable fit in poor-quality panels (such as the untrained panels in Chapter 5 and 6) reflects relative consistency within a weak reference group, not good absolute performance. Fit statistics must therefore be interpreted alongside raw data visualisations (Myford & Wolfe, 2004), particularly when panel baseline performance is inconsistent.

Test design and presentation differences

Differences in test structure between DFC and TIM should be considered when comparing results. The DFC's comparative design with a physical reference present facilitates detection of perceived differences and provides an external anchor for consistent judgment. TIM's monadic presentation requires assessors to rate attributes independently using internal mental references. These design differences may have contributed to differences in separation, reliability, and discrimination, independent of the analytical approaches themselves.

Rating scale design and training

No panel received specific training on the category scale with anchor labels adapted from the LMS. This study did not empirically compare the adapted LMS format against traditional category labels or unstructured line scales. The persistent underuse of the "Strongest imaginable" category across all panels suggests this abstract anchor was problematic regardless of training level. However, whether alternative scale formats would have improved usability, reduced rating variability, or enhanced discrimination cannot be determined without comparative data.

An interesting area for future research would be examining how the MFRM performs with rating scales using relative anchors, commonly employed in QDA settings, and across other scale formats. Meilgaard et al. (2025) emphasise the importance of determining how many scale categories are needed to characterise attribute intensities, which MFRM category diagnostics can empirically evaluate.

Panel quality and diagnostic clarity

MFRM diagnostic value scales with panel quality. The trained panel yielded highly informative diagnostics, while the untrained panels exhibited high variability that obscured patterns and made it difficult to isolate problematic assessors. Even the

selected subset of better-performing assessors in Chapter 6 (n=17 from n=54) displayed residual inconsistency likely due to limited attribute and scale understanding. While the MFRM could potentially identify relatively stable performers within untrained groups for recruitment purposes, some minimum familiarisation with sensory methods and terminology would enhance diagnostic utility.

Experimental design factors also could have influenced panel performance. In Chapter 5, untrained assessors completed evaluations in a single session without rest intervals, likely contributing to fatigue and carryover effects. The absence of incentives may have further reduced engagement. These represent readily addressable constraints in future studies.

7.2.1 Practical implementation considerations

Several practical questions remain unaddressed, including decision rules for fit statistics-based interventions, cost-benefit analysis in industrial settings, and integration with consumer preference data. As discussed in previous chapters, implementation requires specialised software and training in Rasch modelling principles, which may present barriers to adoption. While this thesis demonstrates MFRM's capabilities in research settings, validation within an established practical quality control setting, could demonstrate greater benefits, and remains a valuable direction for future research.

7.3 Knowledge contribution

This thesis provides several contributions to the use of the Many-Facet Rasch Model (MFRM) in sensory difference testing. While previous work has applied MFRM to estimate latent variables such as overall liking or sensory quality, this research extends its application to modelling the overall difference between products based on intensity ratings of multiple sensory attributes. Chapters 4 to 6 demonstrate that the Rasch-derived Total Intensity Measure (TIM) can detect product differences while also indicating which attributes contribute most strongly to those differences and the relative perceptibility of different attributes. This helps address an established limitation in sensory methodology, where existing approaches either focus on single attributes, provide overall metrics without diagnostic detail, or require multiple separate analyses to obtain a full picture.

The research also examines how panel expertise influences the quality and interpretability of MFRM diagnostics. Chapter 5 shows that trained panels produce results that support detailed interpretation of both product and assessor patterns. In contrast, Chapters 4 and 6 show that untrained panels introduce greater variability, which can make some diagnostics more difficult to interpret. An additional observation is that when products differ on opposing attributes and untrained assessors apply ratings inconsistently, the resulting latent difference estimates can show cancellation effects that reduce the apparent magnitude of product differences, with important implications for attribute selection, panel composition, and study design.

A practical framework for applying MFRM in sensory difference testing across different panel types and contexts is also proposed, specifying when and how the method should be applied based on considerations such as panel type, product complexity, and diagnostic needs. By combining Wright maps, category diagnostics, fit statistics, and residual patterns, the framework supports an integrated approach to interpreting product differences, assessor performance, attribute functioning, and scale use. The model's adjustment for systematic differences in assessor severity helps separate genuine performance inconsistencies from individual differences in rating style (i.e., severe or lenient raters).

Additionally, the findings demonstrate how MFRM can provide both measures of overall product difference and insight into the structure and reliability of the underlying data. The ability to trace inconsistent or unexpected findings back to specific assessors or attributes can support more informed interpretation in product development and quality control contexts. By identifying conditions where the method performs well and those where caution is warranted, the thesis offers realistic guidance for integrating MFRM into both research and applied sensory evaluation contexts.

7.4 Recommendations and future perspectives

Building on the findings and discussions presented throughout this thesis, the following recommendations highlight practical improvements and areas for further investigation to enhance Rasch modelling in sensory evaluation.

7.4.1 Review of mean comparisons with Rasch measures

As discussed in Chapter 4, certain mean comparison tests may be unsuitable for analysing Rasch measures, without explicitly modelling interaction effects. The MFRM accounts for main effects across modelled facets and treats unexpected variations as residuals, which are reflected in fit statistics such as OUTFIT Mnsq (Linacre, 1995). Consequently, tests like the Friedman test, which assume interaction structures, may be redundant or misleading in this context.

Instead, the Kruskal-Wallis test was found to be more appropriate for comparing group means on Rasch-transformed measures, as it does not require modelling of interactions and aligns better with the structure of Rasch-derived data. Given these considerations, further research comparing parametric and non-parametric statistical methods for analysing Rasch outputs, could improve methodological transparency, and guide researchers in selecting appropriate tools for post hoc comparisons.

7.4.2 Use of the Partial Credit Rasch Model (PCM) for cross-panel data

In sensory evaluation, it is not uncommon for different panels within the same organisation, often located at geographically distinct sites, to use different rating scales when evaluating the same product. These variations may arise from local practices, differences in panel training, or historical preferences. Such inconsistencies present challenges in data analysis when comparing or consolidating results across panels, as traditional methods typically assume uniform rating structures.

The Partial Credit Model (PCM), described in **Table 2.2. Summary of Rasch Models**, addresses this issue by providing the flexibility to model each attribute using the unique rating scale employed by each panel, accommodating differences in scale structure. Unlike the Rating Scale Model (RSM), which was used in this study and assumes a consistent threshold structure across all attributes and assessors, PCM can accommodate variability in both the number of scale categories and the location of category thresholds. This flexibility makes it particularly well-suited for harmonising data from panels that use different rating formats.

Applying PCM enables calibration of responses from diverse panels onto a common scale, facilitating valid comparisons while preserving the integrity of each panel's original scale. This approach offers a practical method for integrating sensory data

from multiple sources, without enforcing rigid standardisation. It can help balance flexibility with comparability, thereby supporting the generation of more reliable insights to inform product development, quality control, and consumer research.

7.4.3 DIF Analysis for panel proficiency and cross-cultural studies

Monitoring panel performance over time and across different assessor groups is essential for ensuring data quality and maintaining consistent product evaluation standards. Differential Item Functioning (DIF) analysis within the Rasch framework, provides a reliable method for detecting whether specific attributes are interpreted differently by subgroups, even when those groups have similar underlying perceptual sensitivity.

As demonstrated by Myford and Wolfe (2009), Eckes (2023), Shin and Lee (2024), and Lamprianou (2025), MFRM can incorporate time, culture, location, or assessor experience as additional facets. This allows systematic differences to be visualised via Wright maps, offering valuable insights into rater drift and panel dynamics. These capabilities make MFRM particularly useful for both ongoing panel proficiency assessment, and for investigating cultural variability in sensory perception.

In cross-cultural contexts, DIF analysis plays a critical role in validating sensory data from diverse populations. Sensory experiences are shaped by cultural factors, including culinary norms, linguistic framing, and varying familiarity with product categories, which can influence how attributes are perceived and rated (Pangborn et al., 1988; Lee & Lopetcharat, 2017; Hort, 2024; Dupas de Matos et al., 2025). DIF helps uncover latent response biases or semantic mismatches that may arise when comparing panels or test protocols across regions. Persistent DIF in certain attributes may signal a need to adapt scale anchors, redefine terms, or adjust assessor training, to ensure that observed differences reflect genuine sensory perception rather than cultural misalignment.

DIF analysis thus can be a powerful diagnostic tool, enabling researchers and industry practitioners to maintain the integrity of sensory evaluations, while expanding testing across borders or evolving panel compositions. Its application can ensure that decisions derived from sensory data, whether for product reformulation,

quality benchmarking, or market expansion, are based on valid and comparable measurements across global assessor groups.

7.4.4 Application guidance across panel types

The findings across Chapters 4 to 6 demonstrate that the MFRM offers different levels of diagnostic utility depending on panel type and training level. This section provides practical recommendations on when and how to apply the method across different sensory evaluation contexts.

- **Specialist Trained Panels (Industrial QC panels)**

Specialist trained panels represent the optimal context for MFRM application due to their product-specific expertise and consistent exposure, which provide the stable reference framework necessary for informative diagnostics. Chapter 5 demonstrated that trained assessors provided clearer sample discrimination and more stable rating patterns compared to the untrained panel, despite some residual inconsistencies. The MFRM enables ongoing performance monitoring through fit statistics and Wright maps, validates whether new or modified attributes are interpreted consistently by assessors, and adjusts for individual severity differences without requiring identical rating standards.

For multi-site operations where different trained panels use varying rating scales, the Partial Credit Model (section **7.4.2**) enables harmonisation of data across locations by calibrating responses onto a common scale while preserving each panel's original scale structure. DIF analysis (section **7.4.3**) complements this by monitoring panel drift over time and detecting systematic differences between sites, supporting decisions about recalibration needs or protocol standardisation.

These capabilities are particularly valuable for specialist panels, where maintaining consistent standards across sites and over time is critical for quality assurance. Recommended applications include routine quality monitoring, shelf-life testing, reformulation validation, and cross-site comparisons.

- **General Trained (Research) Panels**

Panels with sensory training but less specialised expertise conduct evaluations across different product categories and projects, requiring periodic recalibration. MFRM diagnostics are particularly valuable for these panels in identifying when

recalibration is needed and pinpointing specific assessors or attributes showing drift or inconsistency. The model helps distinguish genuine performance issues from systematic severity differences that do not require intervention.

DIF analysis detects whether subgroups (e.g., assessors with different experience levels or training backgrounds) interpret attributes differently as new products are introduced, informing decisions about where training efforts should focus. Raw data visualisation alongside model outputs is essential for accurate interpretation. Recommended applications include early-stage product formulation testing, comparative testing across product categories, and monitoring performance as panels transition between product types.

- **Untrained and Consumer Panels**

Consumer and untrained panels pose the greatest challenges for product discrimination due to high inter-individual variability and inconsistent scale use. Fit statistics in such groups reflect consistency relative to other poor-performing assessors rather than ideal performance. However, MFRM offers distinctive value by explicitly modelling the separate effects of consumers, products, attributes, and scale steps, producing measurements adjusted for rater severity and erratic scale use.

DIF analysis is particularly valuable for subgroup comparisons, revealing whether attributes or preference patterns function differently across demographic segments, usage occasions, or cultural groups due to culinary norms, linguistic differences, product familiarity, etc. This informs decisions about market segmentation, scale adaptation, or attribute redefinition for global studies. For categorical hedonic or intensity scales commonly used in consumer testing, MFRM's scale diagnostics identify problematic response categories (such as underutilised categories or item polarity i.e. when the interpretation of the scale is reversed), supporting scale optimisation and questionnaire refinement.

Furthermore, the latent variable approach clarifies the relative contribution of each sensory attribute to overall liking or perceived difference, even when consumers cannot articulate these factors explicitly. This supports a measurement-based understanding of consumer drivers of preference rather than relying solely on self-reported reasons. Implementing MFRM effectively with untrained and consumer panels requires careful attribute definition, maintaining identical attribute labels

and scale anchors throughout the study, and sufficient sample sizes to support robust facet estimation and DIF detection.

For semi-trained or familiarised panels with basic orientation to sensory methods, MFRM may additionally serve as a screening tool to identify stable performers for potential recruitment to trained panels, though findings should be interpreted cautiously given the inherent variability in such groups.

- **Academic and Exploratory Panels (Convenience Samples)**

Academic and exploratory research settings often rely on student cohorts, opportunity samples, or mixed consumer participant groups, presenting unique challenges for reliable sensory evaluation. These assessor pools typically lack training, exhibit heterogeneous sensory acuity, and vary widely in task comprehension and familiarity especially with the novel foods and ingredients commonly investigated in academic research settings. Within these contexts, MFRM can explicitly model assessor-related inconsistency, separating it from genuine product differences, and enabling more defensible interpretation of results from these non-standard panels.

DIF analysis is also useful for identifying subgroup patterns, such as differences based on familiarity, dietary orientation, or sensory sensitivity, and for clarifying whether observed product differences are generalisable or segment specific. This is particularly valuable for emerging product categories such as alternative proteins, insect-based foods, novel fermented ingredients, or products using novel processing technologies, where interpretation and acceptance vary widely. Effective application requires clear attribute definitions, justified subgroup classifications, some task familiarisation, and careful attribute selection to minimise construct-irrelevant variance and ensure stable facet estimation.

7.4.5 Software development and usability

This recommendation addresses two areas: improvements to existing Rasch software, such as FACETS, and the integration of Rasch modelling tools into mainstream sensory analysis platforms.

- ***Enhancing the FACETS Software:*** FACETS could be improved by incorporating post hoc sample comparison features directly into the interface, allowing sensory practitioners to streamline analysis and interpretation. Additionally, enhancing

the visual design of Wright maps would make them more suitable for business presentations and stakeholder discussions, increasing their practical impact.

- ***Integration into Sensory Software Platforms:*** While MFRM is powerful, its adoption is limited by the need for specialised software and training. Embedding Rasch modelling functionalities into existing sensory analysis software, supported by simplified user interfaces and clear explanations would significantly broaden access to this method. Although most leading Rasch software packages are paid, open-source implementations are available (Rasch Measurement Transactions, 2025). Additionally, Wind and Hua (2021) and Debelak et al. (2022) offer detailed procedural accounts using R, which can serve as a foundation for integrating Rasch diagnostics into sensory and consumer research programmes.

7.5 Conclusion

This thesis has shown that the Many-Facet Rasch Model (MFRM) can be used to quantify overall product difference as a latent variable while also providing diagnostics for assessors, attributes and scale functioning. Across Chapters 4 to 6, the Total Intensity Measure (TIM) approach detected product differences and offered insight into how panel expertise and attribute structure influence the clarity of the resulting measures. The findings also indicate that inconsistent rating behaviour and opposing attribute profiles can mask genuine product differences through cancellation effects on the latent scale, particularly with variable untrained assessors. The research demonstrated that with adequate panel quality and systematic attribute selection, the Rasch-based approach enables more interpretable comparisons by converting responses to a common interval scale while accounting for systematic assessor differences.

The framework developed in this thesis complements recent applications (Camargo & Henson, 2015b; Ho, 2019; Li, 2019; Chalk, 2020; Mile et al., 2021; Wu et al., 2021) of the MFRM in sensory and consumer research and provides a basis for more informed use of the approach in sensory difference testing. As sensory evaluation continues to involve a wider range of populations and testing contexts, further work in operational quality control settings with trained panels, with alternative scale formats and with more diverse panels, would help clarify the practical boundaries of this approach and strengthen understanding of when it offers the greatest value to sensory practice.

References

Adams, R. J., Wu, M. L., Cloney, D., Berezner, A., & Wilson, M. R. (1997-2020). *ACER ConQuest: Generalised item response modelling software* In Australian Council for Educational Research. https://www.acer.org/au/conquest

Adu, P., Popoola, T., Iqbal, N., Medvedev, O. N., & Simpson, C. R. (2025). Validating the depression anxiety stress scales (DASS-21) across Germany, Ghana, India, and New Zealand using Rasch methodology. *Journal of Affective Disorders*, *383*, 363-373. https://doi.org/10.1016/j.jad.2025.04.099

Aguilera, J. M. (2019). The food matrix: implications in processing, nutrition and health. *Critical Reviews in Food Science and Nutrition*, *59*(22), 3612-3629. https://doi.org/10.1080/10408398.2018.1502743

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%28200002%2980%3A3%3C409%3A%3AAID-JSFA551%3E3.0.CO%3B2-T

Amerine, M. A., Pangborn, R. M., & Roessler, E. B. (1965). *Principles of sensory evaluation of food*. Academic Press. https://doi.org/10.1016/C2013-0-08103-0

Andrés, A. I., Cava, R., Ventanas, J., Thovar, V., & Ruiz, J. (2004). Sensory characteristics of Iberian ham: Influence of salt content and processing conditions. *Meat science*, *68*(1), 45-51. https://doi.org/10.1016/j.meatsci.2003.08.019

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. https://doi.org/10.1007/BF02293814

Andrich, D., & Marais, I. (2019). Equating—Linking Instruments Through Common Items. In *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences* (pp. 137-148). Springer Nature Singapore. https://doi.org/10.1007/978-981-13-7496-8_11

Andrich, D., Sheridan, B., & Luo, G. (1997-2025). *RUMM: Rasch Unidimensional Measurement Models*. In https://www.rummlab.com.au/rumm-2030

Antmann, G., Ares, G., Varela, P., Salvador, A., Coste, B., & Fiszman, S. M. (2011). Consumers' Creaminess Concept Perception: A Cross-Cultural Study in Three Spanish-Speaking Countries. *Journal of Texture Studies*, *42*(1), 50-60. https://doi.org/10.1111/j.1745-4603.2010.00267.x

Arboleda, A. M., Arroyo, C., & Alonso, J. C. (2021). Creating psychometric scales for perceptual assessment of fruit juices' refreshing and thickness attributes. *Appetite*, *163*, 105232. https://doi.org/10.1016/j.appet.2021.105232

Ares, G. (2015). Methodological challenges in sensory characterization. *Current Opinion in Food Science*, *3*, 1-5. https://doi.org/10.1016/j.cofs.2014.09.001

Ares, G. (2018). Methodological issues in cross-cultural sensory and consumer research. *Food Quality and Preference*, *64*, 253-263. https://doi.org/10.1016/j.foodqual.2016.10.007

Ares, G., Barreiro, C., Deliza, R., Giménez, A., & Gámbaro, A. (2010). Application of a Check-All-That-Apply Question to the Development of Chocolate Milk

Desserts. *Journal of Sensory Studies*, *25*(s1), 67-86. https://doi.org/10.1111/j.1745-459X.2010.00290.x

Ares, G., Bruzzone, F., & Giménez, A. (2011). Is A Consumer Panel Able To Reliably Evaluate The Texture Of Dairy Desserts Using Unstructured Intensity Scales? Evaluation Of Global And Individual Performance [Article]. *Journal of Sensory Studies*, *26*(5), 363-370. https://doi.org/10.1111/j.1745-459X.2011.00352.x

Ares, G., Bruzzone, F., Vidal, L., Cadena, R. S., Giménez, A., Pineau, B., Hunter, D. C., Paisley, A. G., & 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

Ares, G., & Varela, P. (2017). Trained vs. consumer panels for analytical testing: Fueling a long lasting debate in the field. *Food Quality and Preference*, *61*, 79-86. https://doi.org/10.1016/j.foodqual.2016.10.006

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

Ariakpomu, N., Ho, P., & Holmes, M. (2025a). *Sensory attribute and difference from control ratings of tomato soup* [Dataset]. https://doi.org/10.5518/1658

Ariakpomu, N. C., Holmes, M., & Ho, P. (2025b). Measuring overall difference from a combination of attribute ratings with the many-facet Rasch model. *Food Quality and Preference*, *127*, 105442. https://doi.org/10.1016/j.foodqual.2025.105442

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Aust, L. B., Gacula, M. C. J., Beard, S. A., & Washam, R. W. I. (1985). Degree of Difference Test Method in Sensory Evaluation of Heterogeneous Product Types. *Journal of Food Science*, *50*(2), 511-513. https://doi.org/10.1111/j.1365-2621.1985.tb13439.x

Bárcenas, P., Elortondo, F. P., & Albisu, M. (2000). Selection and screening of a descriptive panel for ewes milk cheese sensory profiling. *Journal of Sensory Studies*, *15*(1), 79-99. https://doi.org/https://doi.org/10.1111/j.1745-459X.2000.tb00411.x

Barton, A., Hayward, L., Richardson, C. D., & McSweeney, M. B. (2020). Use of different panellists (experienced, trained, consumers and experts) and the projective mapping task to evaluate white wine. *Food Quality and Preference*, *83*, 103900. https://doi.org/10.1016/j.foodqual.2020.103900

Bartoshuk, L. M. (1979). Bitter taste of saccharin related to the genetic ability to taste the bitter substance 6-n-propylthiouracil. *Science*, *205*(4409), 934-935. https://doi.org/10.1126/science.472717

Bartoshuk, L. M., Duffy, V. B., & Miller, I. J. (1994). PTC/PROP tasting: Anatomy, psychophysics, and sex effects. *Physiology & Behavior*, *56*(6), 1165-1171. https://doi.org/10.1016/0031-9384(94)90361-1

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

Bassi, I., Carzedda, M., Gori, E., & Iseppi, L. (2022). Rasch analysis of consumer attitudes towards the mountain product label. *Agricultural and Food Economics*, *10*(1), 13. https://doi.org/10.1186/s40100-022-00218-7

Bechtel, G. G. (1985). Generalizing the Rasch Model for Consumer Rating Scales. *Marketing Science*, *4*(1), 62-73. https://doi.org/10.1287/mksc.4.1.62

Belvedere, S. L., & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of Clinical Epidemiology*, *63*(12), 1287-1297. https://doi.org/10.1016/j.jclinepi.2010.02.012

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bi, J. (2015). *Sensory discrimination tests and measurements: sensometrics in sensory evaluation* (Second ed.). Wiley/Blackwell Publishing. https://doi.org/10.1002/9781118994863

Bi, Y., Zhou, G., Pan, D., Wang, Y., Dang, Y., Liu, J., Jiang, M., & Cao, J. (2019). The effect of coating incorporated with black pepper essential oil on the lipid deterioration and aroma quality of Jinhua ham. *Journal of Food Measurement and Characterization*, *13*(4), 2740-2750. https://doi.org/10.1007/s11694-019-00195-4

Boehm, M. W., Yakubov, G. E., Stokes, J. R., & Baier, S. K. (2020). The role of saliva in oral processing: Reconsidering the breakdown path paradigm. *J Texture Stud*, *51*(1), 67-77. https://doi.org/10.1111/jtxs.12411

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: fundamental measurement in the human sciences* (4th ed.). Routeledge. https://doi.org/10.4324/9780429030499

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 Sci Educ*, *15*(4). https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1), 1416898. https://doi.org/10.1080/2331186X.2017.1416898

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.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. https://pubmed.ncbi.nlm.nih.gov/7154893/

Boutrolle, I., Delarue, J., Arranz, D., Rogeaux, M., & Köster, E. P. (2007). Central location test vs. home use test: Contrasting results depending on product type. *Food Quality and Preference*, *18*(3), 490-499. https://doi.org/10.1016/j.foodqual.2006.06.003

Breslin, P. A. S. (1996). Interactions among salty, sour and bitter compounds. *Trends in Food Science & Technology*, *7*(12), 390-399. https://doi.org/10.1016/s0924-2244(96)10039-x

British Standards Institution. (2019). Sensory analysis. General guidance for the design of test rooms. In. London: BSI Standarads Limited.

British Standards Institution. (2021). Sensory analysis — Methodology — Guidelines for the measurement of the performance of a quantitative descriptive sensory panel. In. London: BSI Standards Limited.

British Standards Institution. (2023). Sensory analysis — Selection and training of sensory assessors. In. London: BSI Standards Limited.

Brockhoff, P. B. (2011). Sensometrics for Food Quality Control. Scandinavian Workshop on Imaging Food Quality 2011: Ystad, May 27, 2011., Technical University of Denmark.

Brockhoff, P. B., Schlich, P., & Skovgaard, I. (2015). Taking individual scaling differences into account by analyzing profile data with the Mixed Assessor Model. *Food Quality and Preference*, *39*, 156-166. https://doi.org/10.1016/j.foodqual.2014.07.005

Brouwer, R., Bouwkamp, T., Scholten, E., Forde, C. G., & Stieger, M. (2024). The effect of viscosity on flavour, mouthfeel and koku enhancement by tastants and yeast extracts in beef broths. *Food Quality and Preference*, *119*, 105235. https://doi.org/10.1016/j.foodqual.2024.105235

Calderón, N., White, B. L., & Seo, H.-S. (2024). Measuring palatability of pet food products: Sensory components, evaluations, challenges, and opportunities. *Journal of Food Science*, *89*(12), 8175-8196. https://doi.org/10.1111/1750-3841.17511

Camargo, F., & Henson, B. (2015a). Aligning Affective Responses with Fabric Features of Vehicle Seat: An Approach Using the Rasch Measurement Model. *International Journal of Affective Engineering*, *14*(3), 193-202. https://doi.org/10.5057/ijae.IJAE-D-15-00012

Camargo, F., & Henson, B. (2015b). Beyond usability: designing for consumers' product experience using the Rasch model. *Journal of Engineering Design*, *26*(4-6), 121-139. https://doi.org/10.1080/09544828.2015.1034254

Castura, J. C., Antúnez, L., Giménez, A., & Ares, G. (2016). Temporal Check-All-That-Apply (TCATA): A novel dynamic method for characterizing products. *Food Quality and Preference*, *47*, 79-90. https://doi.org/10.1016/j.foodqual.2015.06.017

Castura, J. C., Findlay, C. J., & Lesschaeve, I. (2005). Monitoring calibration of descriptive sensory panels using distance from target measurements. *Food Quality and Preference*, *16*(8), 682-690. https://doi.org/10.1016/j.foodqual.2005.03.011

Catley, M. J., O'Connell, N. E., & Moseley, G. L. (2013). How Good Is the Neurophysiology of Pain Questionnaire? A Rasch Analysis of Psychometric Properties. *The Journal of Pain*, *14*(8), 818-827. https://doi.org/10.1016/j.jpain.2013.02.008

Caul, J. F. (1957). The Profile Method of Flavor Analysis. In E. M. Mrak & G. F. Stewart (Eds.), *Advances in Food Research* (Vol. 7, pp. 1-40). Academic Press. https://doi.org/10.1016/S0065-2628(08)60245-1

Cela, N., Condelli, N., Perretti, G., Di Cairano, M., De Clippeleer, J., Galgano, F., & De Rouck, G. (2023). A Comprehensive Comparison of Gluten-Free Brewing

Techniques: Differences in Gluten Reduction Ability, Analytical Attributes, and Hedonic Perception. *Beverages (Basel)*, *9*(1), 18. https://doi.org/10.3390/beverages9010018

Chai, J. J. K., O'Sullivan, C., Gowen, A. A., Rooney, B., & Xu, J.-L. (2022). Augmented/mixed reality technologies for food: A review. *Trends in Food Science & Technology*, *124*, 182-194. https://doi.org/10.1016/j.tifs.2022.04.021

Chalk, C. (2020). The application of Rasch measurement theory to guide consumer product design. Retrieved 20th May 2025, from https://lida.leeds.ac.uk/research-projects/the-application-of-rasch-measurement-theory-to-guide-consumer-product-design/

Chi, S., Liu, X., & Wang, Z. (2021). Comparing student science performance between hands-on and traditional item types: A many-facet Rasch analysis. *Studies in Educational Evaluation*, *70*, 100998. https://doi.org/10.1016/j.stueduc.2021.100998

Chollet, S., Valentin, D., & Abdi, H. (2005). Do trained assessors generalize their knowledge to new stimuli? *Food Quality and Preference*, *16*(1), 13-23. https://doi.org/10.1016/j.foodqual.2003.12.003

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q(3): Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Appl Psychol Meas*, *41*(3), 178-194. https://doi.org/10.1177/0146621616677520

Christensen, K. S., Cosci, F., Carrozzino, D., & Sensky, T. (2024). Rasch Analysis and Its Relevance to Psychosomatic Medicine. *Psychotherapy and Psychosomatics*, *93*(2), 88-93. https://doi.org/10.1159/000535633

Civille, G., & Osdoba, K. (2020). Chapter 4 | Spectrum Descriptive Analysis. In R. N. Bleibaum (Ed.), *Descriptive Analysis Testing for Sensory Evaluation* (Vol. MNL13-2ND-EB, pp. 0). ASTM International. https://doi.org/10.1520/MNL1320150029

Clapham, D., Belissa, E., Inghelbrecht, S., Pensé-Lhéritier, A.-M., Ruiz, F., Sheehan, L., Shine, M., Vallet, T., Walsh, J., & Tuleu, C. (2023). A Guide to Best Practice in Sensory Analysis of Pharmaceutical Formulations. *Pharmaceutics*, *15*(9), 2319. https://doi.org/10.3390/pharmaceutics15092319

Cliff, M. A., Fan, L., Sanford, K., Stanich, K., Doucette, C., & Raymond, N. (2013). Descriptive analysis and early-stage consumer acceptance of yogurts fermented with carrot juice. *Journal of Dairy Science*, *96*(7), 4160-4172. https://doi.org/10.3168/jds.2012-6287

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

Conejo, F., Wooliscroft, B., & Insch, A. (2017). FULL PAPER: Exploring Brand Personality Scale Development Using Rasch Modelling. *Marketing Bulletin*, *27*. https://marketing-bulletin.massey.ac.nz/V27/MB_v27_1_Conejo_2017.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

Corvera-Paredes, B., Sánchez-Reséndiz, A. I., Medina, D. I., Espiricueta-Candelaria, R. S., Serna-Saldívar, S., & Chuck-Hernández, C. (2022). Soft Tribology and Its Relationship With the Sensory Perception in Dairy Products: A Review [Review]. *Frontiers in Nutrition*, *Volume 9 - 2022*. https://doi.org/10.3389/fnut.2022.874763

Cosme, F., Rocha, T., Marques, C., Barroso, J., & Vilela, A. (2025). Innovative Approaches in Sensory Food Science: From Digital Tools to Virtual Reality. *Applied Sciences*, *15*(8), 4538. https://doi.org/10.3390/app15084538

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

Dabb, C., Dryer, R., Brunton, R. J., Krägeloh, C., Moussa, M., Yap, K., Roach, V. J., & Medvedev, O. (2025). Development of the paternal pregnancy-related anxiety scale (PPrAS) using Rasch analysis with Australian and USA samples of expectant fathers. *Journal of Affective Disorders*, *381*, 33-43. https://doi.org/10.1016/j.jad.2025.03.184

De Battisti, F., Nicolini, G., & Salini, S. (2005). The Rasch model to measure the service quality. *The Journal of Services Marketing*, *3*(3), 58-80. https://air.unimi.it/bitstream/2434/6553/2/JCFAI_The%20Rasch%20Model.pdf

Debelak, R., Strobl, C., & Zeigenfuse, M. D. (2022). *An Introduction to the Rasch Model with Examples in R* (1st ed.). CRC Press. https://doi.org/https://doi.org/10.1201/9781315200620

Delarue, J., & Sieffermann, J.-M. (2004). Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food Quality and Preference*, *15*(4), 383-392. https://doi.org/10.1016/S0950-3293(03)00085-5

Deubler, G., Zhang, C., Talavera, M. J., & Swaney-Stueve, M. (2022). Sensory evaluation in the personal care space: A review. *Journal of Sensory Studies*, *37*(6), e12788. https://doi.org/10.1111/joss.12788

Distefano, M., Mauro, R. P., Page, D., Giuffrida, F., Bertin, N., & Leonardi, C. (2022). Aroma Volatiles in Tomato Fruits: The Role of Genetic, Preharvest and Postharvest Factors. *Agronomy*, *12*(2), 376. https://doi.org/10.3390/agronomy12020376

Ditschun, T. L., Riddell, E., Qin, W., Graves, K., Jegede, O., Sharafbafi, N., Pendergast, T., Chidichimo, D., & Wilson, S. F. (2025). Overview of mouthfeel from the perspective of sensory scientists in industry. *Comprehensive Reviews in Food Science and Food Safety*, *24*(2), e70126. https://doi.org/10.1111/1541-4337.70126

Djekic, I., Lorenzo, J. M., Munekata, P. E. S., Gagaoua, M., & Tomasevic, I. (2021). Review on characteristics of trained sensory panels in food science. *Journal of Texture Studies*, *52*(4), 501-509. https://doi.org/10.1111/jtxs.12616

DLG. (2020). *Practice guide for sensory panel training*. https://www.dlg.org/en/mediacenter/dlg-expert-reports/food-sensory-technology

Dong, Y., Sharma, C., Mehta, A., & Torrico, D. D. (2021). Application of Augmented Reality in the Sensory Evaluation of Yogurts. *Fermentation*, *7*(3), 147. https://doi.org/10.3390/fermentation7030147

Doty, R. L., & Cameron, E. L. (2009). Sex differences and reproductive hormone influences on human odor perception. *Physiology & Behavior*, *97*(2), 213-228. https://doi.org/10.1016/j.physbeh.2009.02.032

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241-252. https://doi.org/10.1080/00401706.1964.10490181

Dupas de Matos, A., Chen, A., Maggs, R., Godfrey, A. J. R., Weerawarna N.R.P, M., & Hort, J. (2025). Cross-cultural differences and acculturation in affective response and sensory perception: a case study across Chinese immigrants and local consumers in New Zealand. *Food Quality and Preference*, *122*, 105299. https://doi.org/10.1016/j.foodqual.2024.105299

Eckes, T. (2023). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang. https://doi.org/10.3726/b20875

Engelhard, G. J., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge. https://doi.org/10.4324/9781315766829

Engelhard Jr, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and Composition Program with a Many-Faceted Rasch model. *ETS Research Report Series*, *2003*(1), i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Eskin, D. (2023). Writing Task Performance and First Language Background on an ESL Placement Exam: A Many-Facets Rasch Analysis of Facet Main Effects and Differential Facet Functioning. *Studies in Applied Linguistics & TESOL (SALT)*, *23*(1). https://doi.org/10.52214/salt.v23i1.11805

Fabian, F. W., & Blum, H. B. (1943). Relative Taste Potency of Some Basic Food Constituents and their Competitive and Compensatory Action. *Journal of Food Science*, *8*(3), 179-193. https://doi.org/10.1111/j.1365-2621.1943.tb16560.x

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In *Quantitative Data Analysis for Language Assessment Volume I* (pp. 83-102). Routledge. https://doi.org/10.4324/9781315187815

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

Feeney, E., O'Brien, S., Scannell, A., Markey, A., & Gibney, E. R. (2011). Genetic variation in taste perception: does it have a role in healthy eating? *Proc Nutr Soc*, *70*(1), 135-143. https://doi.org/10.1017/s0029665110003976

Feeney, E. L., McGuinness, L., Hayes, J. E., & Nolden, A. A. (2021). Genetic variation in sensation affects food liking and intake. *Current Opinion in Food Science*, *42*, 203-214. https://doi.org/10.1016/j.cofs.2021.07.001

Ferry, A. L., Hort, J., Mitchell, J. R., Cook, D. J., Lagarrigue, S., & Valles Pamies, B. (2006). Viscosity and flavour perception: Why is starch different from

hydrocolloids? *Food Hydrocolloids*, *20*(6), 855-862. https://doi.org/10.1016/j.foodhyd.2005.08.008

Fidan, D., Çelik Korat, T., Demirgüneş, S., Alaca, S., & Kırömeroğlu, N. (2025). A Tool for Selecting and Writing Texts for Assessing Reading: Reader-Friendly Informative Text Rubric (ODBIMDEPA). *Education and Science*, *50*, 43-68. https://doi.org/10.15390/EB.2025.14166

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.

Freitas, S., Gerardo, P., R., S. M., & Santana, I. (2014). Psychometric Properties of the Montreal Cognitive Assessment (MoCA): An Analysis Using the Rasch Model. *The Clinical Neuropsychologist*, *28*(1), 65-83. https://doi.org/10.1080/13854046.2013.870231

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. https://doi.org/10.1080/01621459.1937.10503522

Frøst, M. B., & Janhøj, T. (2007). Understanding creaminess. *International Dairy Journal*, *17*(11), 1298-1311. https://doi.org/10.1016/j.idairyj.2007.02.007

Fuchs, A., Engleder, S., Huber, J., & Leitner, E. (2022). Principles for the selection of a sensory panel for the evaluation of car interior materials. *Journal of Sensory Studies*, *37*(6), e12782. https://doi.org/10.1111/joss.12782

Fuentes, S., Tongson, E., & Gonzalez Viejo, C. (2021). Novel digital technologies implemented in sensory science and consumer perception. *Current Opinion in Food Science*, *41*, 99-106. https://doi.org/10.1016/j.cofs.2021.03.014

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the Rasch Model. *TPM (Testing Psicometria Metodologia)*, *15*(1), 1-16. https://www.tpmap.org/wp-content/uploads/2014/11/15.1.1.pdf

Garcia, C., Ventanas, J., Antequera, T., Ruiz, J., Cava, R., & Alvarez, P. (1996). Measuring Sensorial Quality of Iberian Ham By Rasch Model. *Journal of Food Quality*, *19*(5), 397-412. https://doi.org/10.1111/j.1745-4557.1996.tb00434.x

Ghalachyan, A., Karpova, E., & Frattali, A. (2024). Developing a holistic sensory evaluation three-part method for textiles and apparel: a practical application for novel materials and products. *Research Journal of Textile and Apparel*, *28*(4), 948-964. https://doi.org/10.1108/RJTA-11-2022-0138

Giacalone, D., & Hedelund, P. I. (2016). Rate-all-that-apply (RATA) with semi-trained assessors: An investigation of the method reproducibility at assessor-, attribute- and panel-level. *Food Quality and Preference*, *51*, 65-71. https://doi.org/10.1016/j.foodqual.2016.02.017

Giezenaar, C., & Hort, J. (2021). A narrative review of the impact of digital immersive technology on affective and sensory responses during product testing in digital eating contexts. *Food Res Int*, *150*(Pt B), 110804. https://doi.org/10.1016/j.foodres.2021.110804

GiftPay. (2024). GiftPay Service providing digital gift cards for use at various retailers in the UK. In *Unified Incentives UK Ltd*: https://www.giftpay.co.uk/.

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

Gilsenan, C. (2010). *An Investigation into Factors Influencing the Sensory Properties of Selected Irish Grown Organic and Conventional Vegetables*. [Doctoral Thesis, Technological University Dublin]. https://arrow.tudublin.ie/tourdoc/21/

González-Pérez, M., Pérez-Garmendia, C., Hoang, K., Susi, R., Antona, B., Barrio, A.-R., & Rosenfield, M. (2025). English version of the Computer Vision Symptom Scale (CVSS17): Translation and Rasch analysis-based cultural adaptation. *PLOS ONE*, *20*(4), e0316936. https://doi.org/10.1371/journal.pone.0316936

Gordon, R. A., Peng, F., Curby, T. W., & Zinsser, K. M. (2021). An introduction to the many-facet Rasch model as a method to improve observational quality measures with an application to measuring the teaching of emotion skills. *Early Childhood Research Quarterly*, *55*, 149-164. https://doi.org/10.1016/j.ecresq.2020.11.005

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. https://doi.org/10.1093/chemse/18.6.683

Grispoldi, L., Zampogni, L., Costanzi, E., Karama, M., El-Ashram, S., Al-Olayan, E., Saraiva, C., García-Díez, J., Iulietto, M. F., & Cenci-Goga, B. (2023). Exploring consumer perception of entomophagy by applying the Rasch model: data from an online survey. *Journal of Insects as Food and Feed*, *10*(1), 9-24. https://doi.org/10.1163/23524588-20230045

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

Großmann, J. L., Westerhuis, J. A., Næs, T., & Smilde, A. K. (2023). Critical evaluation of assessor difference correction approaches in sensory analysis. *Food Quality and Preference*, *106*, 104792. https://doi.org/10.1016/j.foodqual.2022.104792

Guedes, M. D. V., Souza, M. M., Cristini, G. P., Vidor, C. R., & and Kulkamp Guerreiro, I. C. (2021). The use of electronic tongue and sensory panel on taste evaluation of pediatric medicines: a systematic review. *Pharmaceutical Development and Technology*, *26*(2), 119-137. https://doi.org/10.1080/10837450.2020.1860088

Guilleux A, B. M., Hardouin J-B, Sébille V. (2014). Power and Sample Size Determination in the Rasch Model: Evaluation of the Robustness of a Numerical Method to Non-Normality of the Latent Trait. *PLOS ONE*, *9*(1), e83652. https://doi.org/10.1371/journal.pone.0083652

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

Hagell, P. (2014). Testing Rating Scale Unidimensionality Using the Principal Component Analysis (PCA)/t-Test Protocol with the Rasch Model: The Primacy of Theory over Statistics. *Open Journal of Statistics*, *4*(6), 456-465. https://doi.org/10.4236/ojs.2014.46044

Hannum, M., Forzley, S., Popper, R., & Simons, C. T. (2019). Does environment matter? Assessments of wine in traditional booths compared to an immersive and actual wine bar. *Food Quality and Preference*, *76*, 100-108. https://doi.org/10.1016/j.foodqual.2019.04.007

Hariyono, E., Zakhiyah, I., Setiawan, B., Kaniawati, I., & Ishak, A. (2025). Prospective science teachers' knowledge, awareness and understanding of climate change: a case study in Indonesian higher education institution. *International Journal of Sustainability in Higher Education*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/IJSHE-04-2024-0292

Harrington, R. J. (2005). Defining Gastronomic Identity. *Journal of Culinary Science & Technology*, *4*(2-3), 129-152. https://doi.org/10.1300/J385v04n02_10

Hayes, J. E., Allen, A. L., & Bennett, S. M. (2013). Direct comparison of the generalized Visual Analog Scale (gVAS) and general Labeled Magnitude Scale (gLMS). *Food Qual Prefer*, *28*(1), 36-44. https://doi.org/10.1016/j.foodqual.2012.07.012

Heymann, H. (2019). A personal history of sensory science. *Food, Culture & Society*, *22*(2), 203-223. https://doi.org/10.1080/15528014.2019.1573043

Hiğde, E., Yüzüak, A. V., Öcal, Z. M., & Aktamış, H. (2024). A Many-Facet Rasch Measurement Approach to Analyze the Prepared Science Laboratory Activities Based on Science Process Skills and Views of Pre-Service Science Teachers. *Journal of Baltic Science Education*, *23*(4), 641-657. https://doi.org/10.33225/jbse/24.23.641

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. (2015). Chapter 3 - Statistical methods and tools for analysing sensory food texture. In J. Chen & A. Rosenthal (Eds.), *Modifying Food Texture* (pp. 45-87). Woodhead Publishing. https://doi.org/10.1016/B978-1-78242-334-8.00003-1

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. https://doi.org/10.1002/9781119196037

Hollowood, T. A., Linforth, R. S., & Taylor, A. J. (2002). The effect of viscosity on the perception of flavour. *Chem Senses*, *27*(7), 583-591. https://doi.org/10.1093/chemse/27.7.583

Hort, J. (2024). *Cross-Cultural Differences in Perception: A Current Perspective (Keynote Address)*. EuroSense, Dublin.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346-363. https://doi.org/10.1002/bimj.200810425

Huang, Y.-J., Chen, C.-T., Lin, G.-H., Wu, T.-Y., Chen, S.-S., Lin, L.-F., Hou, W.-H., & Hsieh, C.-L. (2018). Evaluating the European Health Literacy Survey Questionnaire in Patients with Stroke: A Latent Trait Analysis Using Rasch

Modeling. *The Patient - Patient-Centered Outcomes Research*, *11*(1), 83-96. https://doi.org/10.1007/s40271-017-0267-3

Ishii, R., Kawaguchi, H., O'Mahony, M., & Rousseau, B. (2007). Relating consumer and trained panels' discriminative sensitivities using vanilla flavored ice cream as a medium. *Food Quality and Preference*, *18*(1), 89-96. https://doi.org/10.1016/j.foodqual.2005.08.004

Issanchou, S. (2015). Sensory & consumer studies with special populations: children and elderly. *Current Opinion in Food Science*, *3*, 53-58. https://doi.org/10.1016/j.cofs.2015.02.004

Jaeger, S. R., Beresford, M. K., Paisley, A. G., Antúnez, L., Vidal, L., Cadena, R. S., Giménez, A., & Ares, G. (2015). Check-all-that-apply (CATA) questions for sensory product characterization by consumers: Investigations into the number of terms used in CATA questions. *Food Quality and Preference*, *42*, 154-164. https://doi.org/10.1016/j.foodqual.2015.02.003

Jaeger, S. R., Meiselman, H. L., & Giacalone, D. (2025). Sensory and consumer science: A complex, expanding, and interdisciplinary field of science. *Food Quality and Preference*, *122*, 105298. https://doi.org/10.1016/j.foodqual.2024.105298

Jotform Inc. (2023). Free online form builder and form creator. In. San Francisco, CA: https://www.jotform.com/.

Kazeniac, S. J., & Hall, R. M. (2006). Flavor chemistry of tomato volatiles. *Journal of Food Science*, *35*(5), 519-530. https://doi.org/10.1111/j.1365-2621.1970.tb04799.x

Keast, R. S. J., & Breslin, P. A. S. (2002). Cross-adaptation and Bitterness Inhibition of L-Tryptophan, L-Phenylalanine and Urea: Further Support for Shared Peripheral Physiology. *Chemical Senses*, *27*(2), 123-131. https://doi.org/10.1093/chemse/27.2.123

Kemp, S. E., Hollowood, T., & Hort, J. (2009). *Sensory Evaluation : A Practical Handbook*. John Wiley & Sons, Incorporated. https://doi.org/10.1002/9781118688076

Kemp, S. E., Hort, J., & Hollowood, T. (2018). *Descriptive Analysis in Sensory Evaluation*. Wiley-Blackwell. https://doi.org/10.1002/9781118991657

Khalaf, M. A., & Omara, E. M. N. (2022). Rasch analysis and differential item functioning of English language anxiety scale (ELAS) across sex in Egyptian context. *BMC Psychology*, *10*(1), 242. https://doi.org/10.1186/s40359-022-00955-w

Kilcast, D., & Clegg, S. (2002). Sensory perception of creaminess and its relationship with food structure. *Food Quality and Preference*, *13*(7), 609-623. https://doi.org/10.1016/S0950-3293(02)00074-5

Kim, D., Kwak, H., Lim, M., & Lee, Y. (2023). Comparison of Check-All-That-Apply (CATA), Rate-All-That-Apply (RATA), Flash Profile, Free Listing, and Conventional Descriptive Analysis for the Sensory Profiling of Sweet Pumpkin Porridge. *Foods*, *12*(19), 3556. https://www.mdpi.com/2304-8158/12/19/3556

Kim, H.-S., Kim, S.-A., & Jang, J.-S. (2025). Validity of the Berg Balance Scale in Individuals With Spinal Cord Injury: A Rasch Analysis. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, *62*, 00469580251338928. https://doi.org/10.1177/00469580251338928

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583-621. https://doi.org/10.1080/01621459.1952.10483441

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean J Anesthesiol*, *70*(2), 144-156. https://doi.org/10.4097/kjae.2017.70.2.144

Lacko, D., Prošek, T., Čeněk, J., Helísková, M., Ugwitz, P., Svoboda, V., Počaji, P., Vais, M., Halířová, H., Juřík, V., & Šašinka, Čeněk. (2023). Analytic and holistic cognitive style as a set of independent manifests: Evidence from a validation study of six measurement instruments. *PLOS ONE*, *18*(6), e0287057. https://doi.org/10.1371/journal.pone.0287057

Lahne, J., & Spackman, C. (2018). Introduction to Accounting for Taste. *The Senses and Society*, *13*(1), 1-5. https://doi.org/10.1080/17458927.2018.1427361

Lamprianou, I. (2025). Network Analysis for the investigation of rater effects in language assessment: A comparison of ChatGPT vs human raters. *Research Methods in Applied Linguistics*, *4*(2), 100205. https://doi.org/10.1016/j.rmal.2025.100205

Lawless, H. T., & Heymann, H. (2010). *Sensory evaluation of foods: principles and practices* (Second, Ed.). Springer. https://doi.org/10.1007/978-1-4419-6488-5

Le Guillas, G., Vanacker, P., Salles, C., & Labouré, H. (2024). Insights to Study, Understand and Manage Extruded Dry Pet Food Palatability. *Animals*, *14*(7), 1095. https://doi.org/10.3390/ani14071095

Lee, H.-S., & Lopetcharat, K. (2017). Effect of culture on sensory and consumer research: Asian perspectives. *Current Opinion in Food Science*, *15*, 22-29. https://doi.org/10.1016/j.cofs.2017.04.003

Lee, Y.-M., Chung, S.-J., Prescott, J., & Kim, K.-O. (2021). Flavor Profiling by Consumers Segmented According to Product Involvement and Food Neophobia. *Foods*, *10*, 598. https://doi.org/10.3390/foods10030598

Lema Almeida, K. A., Koppel, K., & Aldrich, C. G. (2022). Sensory attributes, dog preference ranking, and oxidation rate evaluation of sorghum-based baked treats supplemented with soluble animal proteins. *Journal of Animal Science*, *100*(8). https://doi.org/10.1093/jas/skac191

Lesschaeve, I., Bowen, A., & Bruwer, J. (2012). Determining the impact of consumer characteristics to project sensory preferences in commercial white wines. *American Journal of Enology and Viticulture*, *63*(4), 487-493. https://doi.org/10.5344/ajev.2012.11085

Lestringant, P., Delarue, J., & Heymann, H. (2019). 2010–2015: How have conventional descriptive analysis methods really been used? A systematic review of publications. *Food Quality and Preference*, *71*, 1-7. https://doi.org/10.1016/j.foodqual.2018.05.011

Li, H., Smith, S., Aldrich, G., & Koppel, K. (2018). Preference ranking procedure proposal for dogs: A preliminary study. *Journal of Sensory Studies*, *33*(1), e12307. https://doi.org/10.1111/joss.12307

Li, Z. (2019). *Applications of Rasch analysis in consumer research for new food product development*. [Doctoral Thesis, University of Leeds]. https://etheses.whiterose.ac.uk/24435

Lichters, M., Möslein, R., Sarstedt, M., & Scharf, A. (2021). Segmenting consumers based on sensory acceptance tests in sensory labs, immersive environments, and natural consumption settings. *Food Quality and Preference*, *89*, 104138. https://doi.org/10.1016/j.foodqual.2020.104138

Liem, D. G., Miremadi, F., & Keast, R. S. (2011). Reducing sodium in foods: The effect on flavor. *Nutrients*, *3*(6), 694-711. https://doi.org/10.3390/nu3060694

Lim, J., Wood, A., & Green, B. G. (2009). Derivation and Evaluation of a Labeled Hedonic Scale. *Chemical Senses*, *34*(9), 739-751. https://doi.org/10.1093/chemse/bjp054

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. [PhD Dissertation, The University of Chicago]. https://www.proquest.com/openview/9947a2b1a43f10331c468abcca2ba3e6

Linacre, J. M. (1994). *Many-Facet Rasch measurement*. MESA Press. https://www.winsteps.com/a/Linacre-MFRM-book.pdf

Linacre, J. M. (1995). *ANOVA with Rasch Measures*. American Educational Research Association Annual Meeting, San Francisco, California. https://files.eric.ed.gov/fulltext/ED390942.pdf

Linacre, J. M. (1999). Understanding Rasch Measurement: Estimation Methods for Rasch Measures. *Journal of Outcome Measurements*, *3*(4), 382-405. https://pubmed.ncbi.nlm.nih.gov/10572388/

Linacre, J. M. (2002a). *Guidelines for Rating Scales and Andrich Thresholds*. Retrieved June 19, 2024 from https://rasch.org/rn2.htm

Linacre, J. M. (2002b). Optimizing rating scale category effectiveness. *J Appl Meas*, *3*(1), 85-106. https://pubmed.ncbi.nlm.nih.gov/11997586/

Linacre, J. M. (2004a). From Microscale to Winsteps: 20 years of Rasch software development. *Rasch Measurement Transactions*, *17*(4), 958. https://www.rasch.org/rmt/rmt174g.htm

Linacre, J. M. (2004b). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, *5*(1), 95-110. https://www.winsteps.com/a/Linacre-estimation-further-topics.pdf

Linacre, J. M. (2013). Rasch Measurement Training Seminars: WINSTEPS and Facets (handout). Retrieved 2nd June 2025, from https://winsteps.com/a/handout.pdf

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

Linacre, J. M. (2023a). *Dimensionality: when is a test multidimensional?* Retrieved August 14, 2023 from https://www.winsteps.com/winman/dimensionality.htm

Linacre, J. M. (2023b). *A User's Guide to FACETS Rasch-Model Computer Programs*. https://www.winsteps.com/a/Facets-Manual.pdf

Linacre, J. M. (2024a). *Dimensionality: PCAR contrasts & variances* Retrieved March 16, 2024 from https://www.winsteps.com/winman/principalcomponents.htm

Linacre, J. M. (2024b). *What do Infit and Outfit, Mean-square and Standardized mean?* Retrieved March 14, 2024 from https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2025a). *Facets 64-bit (Many-Facet Rasch Measurement)* In (Version 4.3.0 © 1987-2025) www.winsteps.com

Linacre, J. M. (2025b). *Fit diagnosis: infit outfit mean-square standardized: Winsteps Help*. Retrieved February 18, 2025 from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2025c). *Rater misbehavior*. Retrieved January 15, 2025 from https://www.winsteps.com/facetman64/index.htm?table7agreementstatistics.htm

Linacre, J. M. (2025d). *Table 7 Reliability and Chi-squared Statistics*. Retrieved February 12, 2025 from https://www.winsteps.com/facetman/table7summarystatistics.htm

Linacre, J. M. (2025e). *Winsteps®Rasch Measurement*. In www.winsteps.com

Linacre, M. (2012a). *Many-Facet Rasch Measurement: Facets Tutorial 2 - Fit analysis and Measurement Models* https://www.winsteps.com/a/ftutorial2.pdf

Linacre, M. (2012b). *Many-Facet Rasch Measurement: Facets Tutorial 4 - Anchoring* https://www.winsteps.com/a/ftutorial4.pdf

Liu, J., Liu, M., He, C., Song, H., Guo, J., Wang, Y., Yang, H., & Su, X. (2015). A comparative study of aroma-active compounds between dark and milk chocolate: relationship to sensory perception. *Journal of the Science of Food and Agriculture*, *95*(6), 1362-1372. https://doi.org/10.1002/jsfa.6831

Lu, Y.-M., Wu, Y.-Y., & Lue, Y.-J. (2025). Rasch Analysis of the QuickDASH in Patients with Neck Pain. *Journal of Clinical Medicine*, *14*(6), 1870. https://doi.org/10.3390/jcm14061870

Lubbers, S., Decourcelle, N., Martinez, D., Guichard, E., & Tromelin, A. (2007). Effect of Thickeners on Aroma Compound Behavior in a Model Dairy Gel. *Journal of Agricultural and Food Chemistry*, *55*(12), 4835-4841. https://doi.org/10.1021/jf0628375

Lunz, M. E., & Linacre, J. M. (1998). Measurement Designs Using Multifacet Rasch Modeling. In *Modern Methods for Business Research* (1st ed., pp. 47-77). Taylor and Francis. https://www.taylorfrancis.com/chapters/edit/10.4324/9781410604385-3/

Maheeka, W. N. R. P., Godfrey, A. J. R., Ellis, A., & Hort, J. (2021). Comparing temporal sensory product profile data obtained from expert and consumer panels and evaluating the value of a multiple sip TCATA approach. *Food Quality and Preference*, *89*, 104141. https://doi.org/10.1016/j.foodqual.2020.104141

Mair, P., Hatzinger, R., & Maier, M. J. (2019). *eRm: Extended Rasch Modeling. R package version 1.0-0.* https://CRAN.R-project.org/package=eRm

Majid, A., Burenhult, N., Stensmyr, M., de Valk, J., & Hansson, B. S. (2018). Olfactory language and abstraction across cultures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170139. https://doi.org/doi:10.1098/rstb.2017.0139

Marques, C., Correia, E., Dinis, L.-T., & Vilela, A. (2022). An Overview of Sensory Characterization Techniques: From Classical Descriptive Analysis to the Emergence of Novel Profiling Methods. *Foods*, *11*(3), 255. https://doi.org/10.3390/foods11030255

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

McEwan, J. A., & Lyon, D. H. (2003). Sensory evaluation | Sensory Rating and Scoring Methods. In B. Caballero (Ed.), *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (pp. 5148-5152). Academic Press. https://doi.org/10.1016/B0-12-227055-X/01064-6

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555-576. https://doi.org/10.1177/0265532211430367

Meilgaard, M. C., Civille, G. V., & Carr, B. T. (2015). *Sensory evaluation techniques* (5th ed.). CRC Press. https://doi.org/10.1201/b19493

Meilgaard, M. C., Civille, G. V., Carr, B. T., & Osdoba, K. E. (2025). *Sensory evaluation techniques* (6th ed.). CRC press. https://doi.org/10.1201/9781003352082

Meiselman, H. L. (2013). The future in sensory/consumer research: .....evolving to a better science. *Food Quality and Preference*, *27*(2), 208-214. https://doi.org/10.1016/j.foodqual.2012.03.002

Meiselman, H. L., Jaeger, S. R., Carr, B. T., & Churchill, A. (2022). Approaching 100 years of sensory and consumer science: Developments and ongoing issues [Article]. *Food Quality and Preference*, *100*, Article 104614. https://doi.org/10.1016/j.foodqual.2022.104614

Mello, L. S. S., Almeida, E. L., & Melo, L. (2019). Discrimination of sensory attributes by trained assessors and consumers in semi-sweet hard dough biscuits and their drivers of liking and disliking. *Food Research International*, *122*, 599-609. https://doi.org/10.1016/j.foodres.2019.01.031

Meyners, M., Castura, J. C., & Carr, B. T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference*, *30*(2), 309-319. https://doi.org/10.1016/j.foodqual.2013.06.010

Meyners, M., & Hasted, A. (2021). On the applicability of ANOVA models for CATA data. *Food Quality and Preference*, *92*, 104219. https://doi.org/10.1016/j.foodqual.2021.104219

Microsoft Corporation. (2019). *Microsoft Excel for Microsoft 365 MSO*. In (Version 2408 (Build 16.0.17928.20538) 32-bit)

Mile, L., Sulistijowati, R., Janrianto, W., & Rahman, D. (2021). Acceptability tilapia fish (Oreochromis niloticus) jerky: an application of the many-facets Rasch Model. *IOP Conference Series: Earth and Environmental Science*, *890*(1), 012049. https://doi.org/10.1088/1755-1315/890/1/012049

Miller, K. J., Pollock, C. L., Brouwer, B., & Garland, S. J. (2016). Use of Rasch Analysis to Evaluate and Refine the Community Balance and Mobility Scale for Use in Ambulatory Community-Dwelling Adults Following Stroke. *Phys Ther*, *96*(10), 1648-1657. https://doi.org/10.2522/ptj.20150423

Mitchell, M., Brunton, N. P., & Wilkinson, M. G. (2011). Impact of salt reduction on the instrumental and sensory flavor profile of vegetable soup. *Food Research International*, *44*(4), 1036-1043. https://doi.org/10.1016/j.foodres.2011.03.007

Mohamed-Ahmed, A. H. A., Soto, J., Ernest, T., & Tuleu, C. (2016). Non-human tools for the evaluation of bitter taste in the design and development of medicines: a systematic review. *Drug Discovery Today*, *21*(7), 1170-1180. https://doi.org/10.1016/j.drudis.2016.05.014

Mohsen, T., & Gill, P. (2019). Using the Many-Facet Rasch Model to analyse and evaluate the quality of objective structured clinical examination: a non-

experimental cross-sectional design. *BMJ Open*, *9*(9), e029208. https://doi.org/10.1136/bmjopen-2019-029208

Montero, M. L., & Ross, C. F. (2022). Saltiness perception in white sauce formulations as tested in older adults. *Food Quality and Preference*, *98*, 104529. https://doi.org/10.1016/j.foodqual.2022.104529

Moskowitz, H. R. (2017). Consumers vs experts: Opinions by an outspoken psychophysicist. *Food Quality and Preference*, *61*, 89-91. https://doi.org/10.1016/j.foodqual.2017.01.010

Moskowitz, H. R., & Meiselman, H. L. (2020). *The Origin and Evolution of Human-Centered Food Product Research*. Springer International Publishing. https://doi.org/10.1007/978-3-030-14504-0_161

Muñoz, A. M. (2002). Sensory evaluation in quality control: an overview, new developments and future opportunities. *Food Quality and Preference*, *13*(6), 329-339. https://doi.org/10.1016/S0950-3293(02)00014-9

Muñoz, A. M., Civille, G. V., & Carr, B. T. (1992). *Sensory evaluation in quality control*. Van Nostrand Reinhold. https://doi.org/10.1007/978-1-4899-2653-1

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. https://pubmed.ncbi.nlm.nih.gov/14523257/

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *J Appl Meas*, *5*(2), 189-227. https://pubmed.ncbi.nlm.nih.gov/15064538/

Myford, C. M., & Wolfe, E. W. (2009). Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. *Journal of Educational Measurement*, *46*(4), 371-389. http://www.jstor.org/stable/25651523

Næs, T. (1990). Handling individual differences between assessors in sensory profiling. *Food Quality and Preference*, *2*(3), 187-199. https://doi.org/10.1016/0950-3293(90)90023-N

Næs, T., Brockhoff, P., & Tomic, O. (2010). *Statistics for Sensory And Consumer Science*. https://doi.org/10.1002/9780470669181

Niimi, J., Collier, E. S., Oberrauter, L.-M., Sörensen, V., Norman, C., Normann, A., Bendtsen, M., & Bergman, P. (2022). Sample discrimination through profiling with rate all that apply (RATA) using consumers is similar between home use test (HUT) and central location test (CLT). *Food Quality and Preference*, *95*, 104377. https://doi.org/10.1016/j.foodqual.2021.104377

O'Donnell, K. K. M. W. (2012). Sweeteners and Sugar Alternatives in Food Technology (2nd Edition). In. John Wiley & Sons. https://doi.org/10.1002/9781118373941

Oon, P.-T., & and Fan, X. (2017). Rasch analysis for psychometric improvement of science attitude rating scales. *International Journal of Science Education*, *39*(6), 683-700. https://doi.org/10.1080/09500693.2017.1299951

Othman, A., Saad, N., & Rani, A. M. A. (2021). Quantifying Haptic Qualities for Real World Automotive Applications. Proceedings of the 19th annual conference of Asia Digital Art and Design Association-ADADA+ CUMULUS 2021, Korea.

Owusu, G., Sylvester, J., & Edi, D. H. (2022). Influence of Types of Soymilk Formulations on Sensory Characteristics: A Study in Northern Region of

Ghana. *Asian Journal of Science and Technology*, *13*(1), 12020-12024. https://www.researchgate.net/publication/358273884

Pagani, L., & Zanarotti, M. C. (2010). Some Uses of Rasch Models Parameters in Customer Satisfaction Data Analysis. *Quality Technology & Quantitative Management*, *7*(1), 83-95. https://doi.org/10.1080/16843703.2010.11673220

Pagès, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference*, *16*(7), 642-649. https://doi.org/10.1016/j.foodqual.2005.01.006

Pagès, J., Cadoret, M., & Lê, S. (2010). The Sorted Napping: A New Holistic Approach in Sensory Evaluation. *Journal of Sensory Studies*, *25*(5), 637-658. https://doi.org/10.1111/j.1745-459X.2010.00292.x

Pangborn, R. M., Guinard, J.-X., & Davis, R. G. (1988). Regional aroma preferences. *Food Quality and Preference*, *1*(1), 11-19. https://doi.org/10.1016/0950-3293(88)90003-1

Parker, J. K., Methven, L., Pellegrino, R., Smith, B. C., Gane, S., & Kelly, C. E. (2022). Emerging Pattern of Post-COVID-19 Parosmia and Its Effect on Food Perception. *Foods*, *11*(7), 967. https://doi.org/10.3390/foods11070967

Peltier, C., Mammasse, N., Visalli, M., Cordelle, S., & Schlich, P. (2018). Do we need to replicate in sensory profiling studies? *Food Quality and Preference*, *63*, 129-134. https://doi.org/10.1016/j.foodqual.2017.09.001

Peryam, D. R., & Pilgrim, F. J. (1957). Hedonic scale method of measuring food preferences. *Food technology*.

Pham, A., Schilling, M., Mikel, W., Williams, J., Martin, J., & Coggins, P. (2008). Relationships between sensory descriptors, consumer acceptability and volatile flavor compounds of American dry-cured ham. *Meat science*, *80*(3), 728-737. https://doi.org/10.1016/j.meatsci.2008.03.015

Pineau, N., Girardi, A., Lacoste Gregorutti, C., Fillion, L., & Labbe, D. (2022). Comparison of RATA, CATA, sorting and Napping® as rapid alternatives to sensory profiling in a food industry environment. *Food Research International*, *158*, 111467. https://doi.org/10.1016/j.foodres.2022.111467

Pineau, N., Schlich, P., Cordelle, S., Mathonnière, C., Issanchou, S., Imbert, A., Rogeaux, M., Etiévant, P., & Köster, E. (2009). Temporal Dominance of Sensations: Construction of the TDS curves and comparison with time–intensity. *Food Quality and Preference*, *20*(6), 450-455. https://doi.org/10.1016/j.foodqual.2009.04.005

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

Poirson, E., Petiot, J.-F., & Richard, F. (2010). A method for perceptual evaluation of products by naive subjects: Application to car engine sounds. *International Journal of Industrial Ergonomics*, *40*(5), 504-516. https://doi.org/10.1016/j.ergon.2010.06.001

Polat, M. (2020). A Rasch Analysis of Rater Behaviour in Speaking Assessment. *International Online Journal of Education and Teaching*, *7*(3), 1126-1141. https://eric.ed.gov/?id=EJ1258433

Politi, M. T., Ferreira, J. C., & Patino, C. M. (2021). Nonparametric statistical tests: friend or foe? *Jornal Brasileiro de Pneumologia*, *47*(4), e20210292. https://doi.org/10.36416/1806-3756/e20210292

Posit Team. (2023). *RStudio: Integrated Development Environment for R*. In (Version 2023.03.1+446 "Cherry Blossom") Posit Software, PBC. https://www.posit.co/.

Prasetyaningrum, P. T., Purwanto, P., & Rochim, A. F. (2024). Analyzing Customer Engagement with Gamification Approach in the Banking Sector Using the Rasch Model. *Ingénierie des Systèmes d'Information*, *29*(3). https://www.iieta.org/journals/isi/paper/10.18280/isi.290323

Puputti, S., Aisala, H., Hoppu, U., & Sandell, M. (2019). Factors explaining individual differences in taste sensitivity and taste modality recognition among Finnish adults. *Journal of Sensory Studies*, *34*(4). https://doi.org/10.1111/joss.12506

Quansah, F. (2022). Item and rater variabilities in students' evaluation of teaching in a university in Ghana: application of Many-Facet Rasch Model. *Heliyon*, *8*(12), e12548. https://doi.org/10.1016/j.heliyon.2022.e12548

Raithatha, C., & Rogers, L. (2018). Chapter 4 - Panel Quality Management: Performance, Monitoring and Proficiency. In S. E. Kemp, J. Hort, & T. Hollowood (Eds.), *Descriptive Analysis in Sensory Evaluation*. Wiley-Blackwell. https://doi.org/10.1002/9781118991657.ch4

Ramp, M., Khan, F., Misajon, R. A., & Pallant, J. F. (2009). Rasch analysis of the Multiple Sclerosis Impact Scale MSIS-29. *Health Qual Life Outcomes*, *7*, 58. https://doi.org/10.1186/1477-7525-7-58

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research (Republished by University of Chicago Press, 1980). https://archive.org/details/probabilisticmod0000rasc

Rasch Measurement Transactions. (2025). *Rasch Measurement Analysis Software Directory*. https://www.rasch.org/software.htm

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

Robitzsch, A., Kiefer, T., & Wu, M. (2021). *TAM: Test Analysis Modules. R package version 3.6-4*. https://CRAN.R-project.org/package=TAM

Rogers, L. (2017). *Discrimination testing in Sensory science: A practical handbook* (First ed.). Woodhead Publishing. https://www.sciencedirect.com/book/9780081010099

Rogues, J., Csoltova, E., Larose-Forges, C., & Mehinagic, E. (2022). 16 - Sensory evaluation of pet food products. In A.-M. Pensé-Lhéritier, I. Bacle, & J. Delarue (Eds.), *Nonfood Sensory Practices* (pp. 313-329). Woodhead Publishing. https://doi.org/10.1016/B978-0-12-821939-3.00011-7

Romano, R., Brockhoff, P. B., Hersleth, M., Tomic, O., & Næs, T. (2008). Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis. *Food Quality and Preference*, *19*(2), 197-209. https://doi.org/10.1016/j.foodqual.2007.06.008

Ruiz-Capillas, C., & Herrero, A. M. (2021). Sensory Analysis and Consumer Research in New Product Development. *Foods*, *10*(3), 582. https://doi.org/10.3390/foods10030582

Saint-Eve, A., Kora, E. P., & Martin, N. (2004). Impact of the olfactory quality and chemical complexity of the flavouring agent on the texture of low fat stirred yogurts assessed by three different sensory methodologies. *Food Quality and Preference*, *15*(7-8), 655-668. https://doi.org/10.1016/j.foodqual.2003.09.002

Salzberger, T., & Sinkovics, R. R. (2006). Reconsidering the problem of data equivalence in international marketing research. *International Marketing Review*, *23*(4), 390-417. https://doi.org/10.1108/02651330610678976

Samir, A., & Tabatabaee-Yazdi, M. (2020). Translation Quality Assessment Rubric: A Rasch Model-Based Validation. *International Journal of Language Testing*, *10*(2), 101-128. https://eric.ed.gov/?id=EJ1291125

Sanderson, T., & Hollowood, T. (2017). Temporal Methods for Assessment of Household and Personal Care Products. In *Time-Dependent Measures of Perception in Sensory Evaluation* (pp. 362-387). https://doi.org/10.1002/9781118991640.ch14

Schifferstein, H. N. J. (2012). Labeled magnitude scales: A critical review. *Food Quality and Preference*, *26*(2), 151-158. https://doi.org/10.1016/j.foodqual.2012.04.016

Schlossareck, C., & Ross, C. F. (2019). Electronic Tongue and Consumer Sensory Evaluation of Spicy Paneer Cheese. *Journal of Food Science*, *84*(6), 1563-1569. https://doi.org/10.1111/1750-3841.14604

Shapin, S. (2016). A taste of science: Making the subjective objective in the California wine world. *Social Studies of Science*, *46*(3), 436–460. http://www.jstor.org/stable/26099849

Sharma, C. (2023). Language of smell: Tracing some cross-cultural insights from past and present [Review]. *Frontiers in Food Science and Technology*, *Volume 3 - 2023*. https://doi.org/10.3389/frfst.2023.1091355

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies*, *29*, 24735-24757. https://doi.org/10.1007/s10639-024-12817-6

Sick, J. (2010). Rasch measurement in language education - part 5. *SHIKEN: JALT Testing and evaluation SIG Newsletter*, *14*(2), 23-29. https://teval.jalt.org/test/sic_1.htm

Sinesio, F., Moneta, E., Porcherot, C., Abbà, S., Dreyfuss, L., Guillamet, K., Bruyninckx, S., Laporte, C., Henneberg, S., & McEwan, J. A. (2019). Do immersive techniques help to capture consumer reality? *Food Quality and Preference*, *77*, 123-134. https://doi.org/10.1016/j.foodqual.2019.05.004

Sipos, L., Ágoston, K. C., Biró, P., Bozóki, S., & Csató, L. (2025). How to measure consumer's inconsistency in sensory testing? *Curr Res Food Sci*, *10*, 100982. https://doi.org/10.1016/j.crfs.2025.100982

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

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals.

Journal of Applied Measurement, *3*(2), 205-231. https://pubmed.ncbi.nlm.nih.gov/12011501/

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*(2), 199-218. https://pubmed.ncbi.nlm.nih.gov/12029178/

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680. https://doi.org/10.1126/science.103.2684.677

Stokes, J. R., Boehm, M. W., & Baier, S. K. (2013). Oral processing, texture and mouthfeel: From rheology to tribology and beyond. *Current Opinion in Colloid & Interface Science*, *18*(4), 349-359. https://doi.org/10.1016/j.cocis.2013.04.010

Stolt, M., Kottorp, A., & Suhonen, R. (2022). The use and quality of reporting of Rasch analysis in nursing research: A methodological scoping review. *International Journal of Nursing Studies*, *132*, 104244. https://doi.org/10.1016/j.ijnurstu.2022.104244

Stone, H., Bleibaum, R. N., & Thomas, H. A. (2012). *Sensory evaluation practices* (4th ed.). Academic. https://www.sciencedirect.com/book/9780123820860/sensory-evaluation-practices

Stone, H., Bleibaum, R. N., & Thomas, H. A. (2020). *Sensory evaluation practices* (5th ed.). Academic press. https://doi.org/10.1016/C2017-0-03038-0

Stone, H., Sidel, J., Oliver, S., Woolsey, A., & Singleton, R. C. (2004). Sensory Evaluation by Quantitative Descriptive Analysis. In *Descriptive Sensory Analysis in Practice* (pp. 23-34). https://doi.org/10.1002/9780470385036.ch1c

Tarricone, P., & Cooper, M. G. (2014). Using Rasch measurement to improve analytical marking keys. *Assessment Matters*, *6*, 86-111. https://doi.org/10.18296/am.0118

Tavakol, M., & and Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teacher*, *35*(1), e838-e848. https://doi.org/10.3109/0142159X.2012.737488

Taylor, W. J., & McPherson, K. M. (2007). Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. *Arthritis Care & Research*, *57*(5), 723-729. https://doi.org/10.1002/art.22770

Teillet, E., Schlich, P., Urbano, C., Cordelle, S., & Guichard, E. (2010). Sensory methodologies and the taste of water. *Food Quality and Preference*, *21*(8), 967-976. https://doi.org/10.1016/j.foodqual.2010.04.012

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum*, *57*(8), 1358-1362. https://doi.org/10.1002/art.23108

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch Analysis in the Development and Application of Quality of Life Instruments. *Value in Health*, *7*(s1), S22-S26. https://doi.org/10.1111/j.1524-4733.2004.7s106.x

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 2. Advanced model applications and the data-

model fit assessment. *Disability and Rehabilitation*, *46*(3), 604-617. https://doi.org/10.1080/09638288.2023.2169772

The John Hopkins University Hospital. (2023). *Smell and taste disorders*. The Johns Hopkins Hospital, and Johns Hopkins Health System. Retrieved August 27, 2023 from https://www.hopkinsmedicine.org/health/conditions-and-diseases/smell-and-taste-disorders

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

Thurstone, L. L. (1927). A law of comparative jugdement. *Psychological Review*, *34*, 273. https://doi.org/10.1037/h0070288

Tomaschunas, M., Köhn, E., Bennwitz, P., Hinrichs, J., & Busch-Stockfisch, M. (2013). Quantitative and Qualitative Variation of Fat in Model Vanilla Custard Desserts: Effects on Sensory Properties and Consumer Acceptance. *Journal of Food Science*, *78*(6), S894-S901. https://doi.org/10.1111/1750-3841.12128

Tomic, O., Luciano, G., Nilsen, A., Hyldig, G., Lorensen, K., & Næs, T. (2010). Analysing sensory panel performance in a proficiency test using the PanelCheck software. *European Food Research and Technology*, *230*(3), 497-511. https://doi.org/10.1007/s00217-009-1185-y

Tomic, O., Nilsen, A., Martens, M., & Næs, T. (2007). Visualization of sensory profiling data for performance monitoring. *LWT - Food Science and Technology*, *40*(2), 262-269. https://doi.org/10.1016/j.lwt.2005.09.014

Torrico, D. D., Han, Y., Sharma, C., Fuentes, S., Gonzalez Viejo, C., & Dunshea, F. R. (2020). Effects of Context and Virtual Reality Environments on the Wine Tasting Experience, Acceptability, and Emotional Responses of Consumers. *Foods*, *9*(2), 191. https://doi.org/10.3390/foods9020191

Torrico, D. D., Mehta, A., & Borssato, A. B. (2023). New methods to assess sensory responses: a brief review of innovative techniques in sensory evaluation. *Current Opinion in Food Science*, *49*, 100978. https://doi.org/10.1016/j.cofs.2022.100978

Tournier, C., Sulmont-Rossé, C., & Guichard, E. (2007). Flavour perception: Aroma, taste and texture interactions. *Food*, *1*(2), 246-257. https://www.researchgate.net/publication/284561985

Touzani, R., Rouquette, A., Schultz, E., Allaire, C., Carrieri, P., Mancini, J., & Hardouin, J.-B. (2024). Psychometric validation of the French version of two scales measuring general (HLS19-Q12) and navigational (HLS19-NAV) health literacy using the Rasch model. *BMC Public Health*, *24*(1), 3079. https://doi.org/10.1186/s12889-024-20504-x

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99-114. https://doi.org/10.2307/3001913

Turek, P., & Kowalska, M. (2024). The use of model samples in the process of selection of sensory panel to assess cosmetic products. *Journal of Sensory Studies*, *39*(1), e12895. https://doi.org/10.1111/joss.12895

Üren, N. (2024). Constructing a descriptive sensory panel for tactile comfort evaluations: Effect of demographic variables and panel size. *International Advanced Researches and Engineering Journal*, *8*(1), 51-60. https://doi.org/10.35860/iarej.1380044

van Zyl, H., & Meiselman, H. L. (2016). An update on the roles of culture and language in designing emotion lists: English, Spanish and Portuguese. *Food*

Quality and Preference, *51*, 72-76. https://doi.org/10.1016/j.foodqual.2016.02.019

Veldhuizen, M. G., Siddique, A., Rosenthal, S., & Marks, L. E. (2017). Interactions of Lemon, Sucrose and Citric Acid in Enhancing Citrus, Sweet and Sour Flavors. *Chemical Senses*, *43*(1), 17-26. https://doi.org/10.1093/chemse/bjx063

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. https://doi.org/10.1007/978-0-387-21706-2

Verriele, M., Plaisance, H., Vandenbilcke, V., Locoge, N., Jaubert, J. N., & Meunier, G. (2012). Odor Evaluation and Discrimination of Car Cabin And its Components: Application of the "Field of Odors" Approach in a Sensory Descriptive Analysis. *Journal of Sensory Studies*, *27*(2), 102-110. https://doi.org/10.1111/j.1745-459X.2012.00371.x

Vidal, L., Ares, G., Hedderley, D. I., Meyners, M., & Jaeger, S. R. (2018). Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies. *Food Quality and Preference*, *67*, 49-58. https://doi.org/10.1016/j.foodqual.2016.12.013

Wang, D., Wang, J., Lang, Y., Huang, M., Hu, S., Liu, H., Sun, B., Long, Y., Wu, J., & Dong, W. (2025). Interactions between food matrices and odorants: A review. *Food Chemistry*, *466*, 142086. https://doi.org/10.1016/j.foodchem.2024.142086

Wang, H., Feng, X., Suo, H., Yuan, X., Zhou, S., Ren, H., Jiang, Y., & Kan, J. (2022). Comparison of the performance of the same panel with different training levels: Flash profile versus descriptive analysis. *Food Quality and Preference*, *99*, 104582. https://doi.org/10.1016/j.foodqual.2022.104582

Wang, S., Olarte Mantilla, S. M., Smith, P. A., Stokes, J. R., & Smyth, H. E. (2020). Astringency sub-qualities drying and pucker are driven by tannin and pH – Insights from sensory and tribology of a model wine system. *Food Hydrocolloids*, *109*, 106109. https://doi.org/10.1016/j.foodhyd.2020.106109

Wang, S., Olarte Mantilla, S. M., Smith, P. A., Stokes, J. R., & Smyth, H. E. (2021). Tribology and QCM-D approaches provide mechanistic insights into red wine mouthfeel, astringency sub-qualities and the role of saliva. *Food Hydrocolloids*, *120*, 106918. https://doi.org/10.1016/j.foodhyd.2021.106918

Whelan, V. J. (2017). Chapter 11 - Difference From Control (DFC) Test. In L. Rogers (Ed.), *Discrimination Testing in Sensory Science* (pp. 209-236). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-101009-9.00011-3

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wind, S., & Hua, C. (2021). *Rasch Measurement Theory Analysis in R: Illustrations and Practical Guidance for Researchers and Practitioners*. Bookdown. https://bookdown.org/chua/new_rasch_demo2/

Worch, T., Lê, S., & Punter, P. (2010). How reliable are the consumers? Comparison of sensory profiles from consumers and experts. *Food Quality and Preference*, *21*(3), 309-318. https://doi.org/10.1016/j.foodqual.2009.06.001

Worch, T., Sinesio, F., Moneta, E., Abbà, S., Dreyfuss, L., McEwan, J. A., & Porcherot-Lassallette, C. (2020). Influence of different test conditions on

the emotional responses elicited by beers. *Food Quality and Preference*, *83*, 103895. https://doi.org/10.1016/j.foodqual.2020.103895

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. . *Educational and Psychological Measurement*, *29*(1), 23-48. https://doi.org/10.1177/001316446902900102

Wright, B. D. (1991). Scores, reliabilities and assumptions. *Rasch Measurement Transactions*, *5*(3), 157-158. https://www.rasch.org/rmt/rmt53a.htm

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. https://research.acer.edu.au/measurement/2/

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

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339-355. https://pubmed.ncbi.nlm.nih.gov/24064576/

Wu, M., Gani, H., Viney, S., Ho, P., & Orfila, C. (2021). Effect of ginger-enriched pasta on acceptability and satiety. *International Journal of Food Science and Technology*, *56*(9), 4604-4614. https://doi.org/10.1111/ijfs.15264

Xiangli, R., Ma, Y., Zeng, Y., Tang, K., Chen, S., & Xu, Y. (2024). Differences and correlations between industrial experts and semi-trained assessors in the sensory evaluation of strong-aroma baijiu using rate-all-that-apply. *Journal of Food Science*, *89*(9), 5841-5857. https://doi.org/10.1111/1750-3841.17280

Yang, J., & Lee, J. (2019). Application of Sensory Descriptive Analysis and Consumer Studies to Investigate Traditional and Authentic Foods: A Review. *Foods*, *8*(2), 54. https://doi.org/10.3390/foods8020054

Yang, Q., Nijman, M., Flintham, M., Tennent, P., Hidrio, C., & Ford, R. (2022). Improving simulated consumption context with virtual Reality: A focus on participant experience. *Food Quality and Preference*, *98*, 104531. https://doi.org/10.1016/j.foodqual.2022.104531

Yao, E., Lim, J., Tamaki, K., Ishii, R., Kwang-Ok, & O'Mahony, M. (2003). Structured and unstructured 9-point hedonic scales: A cross cultural study with American, Japanese and Korean consumers. *Journal of Sensory Studies*, *18*(2), 115-139. https://doi.org/10.1111/j.1745-459X.2003.tb00379.x

Zeppa, G., Bertolino, M., & Rolle, L. (2012). Quantitative descriptive analysis of Italian polenta produced with different corn cultivars. *Journal of the Science of Food and Agriculture*, *92*(2), 412-417. https://doi.org/10.1002/jsfa.4593

Zhang, Z. (1996). Teacher Assessment Competency: A Rasch Model Analysis. https://eric.ed.gov/?id=ED400322

Zhi, R., Zhao, L., & Shi, J. (2016). Improving the sensory quality of flavored liquid milk by engaging sensory analysis and consumer preference. *Journal of Dairy Science*, *99*(7), 5305-5317. https://doi.org/10.3168/jds.2015-10612

Appendices

Appendix A : Ethics Approval Letters

A.1 Jaffa cakes study ethics approval (AREA FREC 2023-0433-496)

13 April 2023

Dear Nnenna Ariakpomu,

0433 - Application of Rasch Analysis in Sensory Difference Testing

NB: All approvals/comments are subject to compliance with current University of Leeds and UK Government advice regarding the Covid-19 pandemic.

I am pleased to inform you that the above research ethics application has been reviewed by the Business, Earth & Environment, Social Sciences (AREA FREC) Committee and on behalf of the Chair, I can confirm a favourable ethical opinion based on the documentation received at date of this email.

Please retain this email as evidence of approval in your study file.

Please notify the committee if you intend to make any amendments to the original research as submitted and approved to date. This includes recruitment methodology; all changes must receive ethical approval prior to implementation. Please see https://ris.leeds.ac.uk/research-ethics-and-integrity/applying-for-an-amendment/ or contact the Research Ethics Administrator for further information (EthicsEnquiries@leeds.ac.uk) if required.

Ethics approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, risk assessments and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

It is our policy to remind everyone that it is your responsibility to comply with Health and Safety, Data Protection and any other legal and/or professional guidelines there may be.

I hope the study goes well.

Best wishes

Ms Rachel De Souza, Lead Research Ethics & Governance Administrator, Secretariat

On behalf of Dr Judith Hanks, Chair, Joint AREA FREC

Page 1 of 1

A.2 Chocolate spreads study ethics approval (MEEC 15 -003)

The Secretariat
Level 11, Worsley Building
University of Leeds
Leeds, LS2 9NL
Tel: 0113 343 4873
Email: ResearchEthics@leeds.ac.uk

UNIVERSITY OF LEEDS

Dr Peter Ho
School of Food Science and Nutrition
University of Leeds
Leeds, LS2 9JT

MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC)
University of Leeds

4 May 2017

Dear Peter

Title of study	**Introduction to Food Product Development**
Ethics reference	**MEEC 15-003 amendment Oct 2016**
Module references	**FOOD2192; FOOD3371; FOOD5472M; FOOD5455M;FOOD3050; FOOD5071M**

I am pleased to inform you that the amendment to the application listed above has been reviewed by a representative of the MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC) and I can confirm a favourable ethical opinion as of the date of this letter. The following documentation was considered:

Document	*Version*	*Date*
MEEC 15-003 amendment May 2017 Amendmentv3_form MEEC_003.doc	1	04/05/17
MEEC 15-003 amendment May 2017 Application_form_for_taught_student_modules__block_approval_food update2017.doc	1	04/05/17
MEEC 15-003 amendment Oct 2016 Amendment_form MEEC_003.doc	1	05/10/16
MEEC 15-003 amendment Oct 2016 Application_form_for_taught_student_modules__block_approval_food update.doc	1	05/10/16
MEEC 15-003 Application_form_for_taught_student_modules__block_approval_food.doc	2	09/10/15
MEEC 15-003 Focus group Consent Form food.docx	2	09/10/15
MEEC 15-003 Survey Consent Form food.docx	2	09/10/15
MEEC 15-003 Taste panel practical Consent Form food.docx	2	09/10/15
MEEC 15-003 Taste panel project Consent Form food.docx	2	09/10/15

Please notify the committee if you intend to make any further amendments to the original research as submitted at date of this approval. All changes must receive ethical approval prior to implementation. The amendment form is available at http://ris.leeds.ac.uk/EthicsAmendment.

Please note: You will be given a two week notice period if your project is to be audited. There is a checklist listing examples of documents to be kept which is available at http://ris.leeds.ac.uk/EthicsAudits.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to ResearchEthics@leeds.ac.uk.

Yours sincerely

Jennifer Blaikie
Senior Research Ethics Administrator, Research & Innovation Service
On behalf of Dr Dawn Groves, Chair, MEEC FREC

A.3 Tomato soup study ethics approval (BESS+FREC - 2024 0433-2568)

UNIVERSITY OF LEEDS

4 November 2024

Dear Nnenna,

Your research ethics application reference: 0433

Amendment reference number: BESS+ FREC - 2024 0433-2568

Your research project: Application of Rasch Analysis in Sensory Difference Testing

I am pleased to inform you that the above amendment application has been reviewed by the Business, Environment, Social Sciences (BESS+ FREC) Faculty Research Ethics Committee (FREC) which has issued a favourable ethical opinion based on the application submitted. **Please retain this email in your project file as it is evidence of the Committee's approval.**

Matters you should note:

- Ethics approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The Committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.
- It is your responsibility to comply with all relevant Health and Safety, Data Protection and other legal and professional requirements and guidelines.
- You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, risk assessments and other documents relating to the research project. This should be kept in your project file.
- Audits are undertaken on approved ethics applications. Your project could be chosen for such an audit. You should therefore ensure your project files are kept up to date and readily available for audit purposes. You will be given a two week notice period if your project is selected.
- Please always include the above research ethics application reference and Amendment request reference in any correspondence with the Research Ethics team.

If you need to make amendments to the original research project as submitted, you are expected to seek approval from the Committee before taking any further action. Changes could include (but are not limited to) the project end date, project design or recruitment methodology, or study documentation. Please go to https://secretariat.leeds.ac.uk/research-ethics/how-to-apply-for-research-ethics-amendment/ or contact the Research Ethics team for further information at Research Ethics.

I hope your research project continues to go well.

Yours sincerely,

Ms Taylor Haworth, Phoenix Lead, Research Ethics, Secretariat, University of Leeds

On behalf of Dr Judith Hanks, Chair, BESS+ FREC

Appendix B : Composition of Samples

B.1 Jaffa Cakes Sample Content

Table B 1. Sample composition for the three Jaffa cake samples in the Chapter 4 study

Attributes of Interest	*Content per 100g*	Brand A	Brand B	Brand C/Reference
	Energy (Kcal)	377	385	388
Orange Flavour	Total Fat (g)	8.1	9	9.4
Sweetness	Saturates(g)	4	4.8	5.1
Cocoa Flavour	Total Carbs (g)	69.8	70	70
Milky Flavour	Sugars (g)	50.2	46	51.5
Saltiness	Fibre (g)	2.1	2.1	2.2
	Protein (g)	5	4.6	4.7
	Salt (g)	0.27	0.2	1.9
	ORANGE FLAVOURING			
	Concentrated Orange Juice	✗	1%	1%
Comparison of Orange, Cocoa, and Milk flavour Content	Orange Juice Equivalent	8%	✗	✗
	COCOA CONTENT			
	Dark Chocolate	19%	22%	22%
	ADDED MILK CONTENT			
	Milk	as Butter oil	✗	✗
List of Ingredients		Glucose-Fructose Syrup, Dark Chocolate (19%) [Sugar, Cocoa Mass, Vegetable Fats (Palm, Shea), Butter Oil (Milk), Cocoa Butter, Emulsifiers (Soya Lecithin, E476), Natural Flavouring], Sugar, Flour (Wheat Flour, Calcium, Iron, Niacin, Thiamin), ***Whole Egg***, Water, Dextrose, Concentrated Orange Juice, Glucose Syrup, Vegetable Oils (Sunflower, Palm), Humectant (Glycerine), Gelling Agent (Pectin), Acid (Citric Acid), Raising Agents (Ammonium Bicarbonate, Disodium Diphosphate, Sodium Bicarbonate), ***Dried Whole Egg***, Acidity Regulator (Sodium Citrates), Natural Orange Flavouring, Colour (Curcumin), Emulsifier (Soya Lecithin), Product contains the equivalent of 8% Orange Juice	Glucose-Fructose Syrup, Dark Chocolate (22%) [Sugar, Cocoa Mass, Cocoa Butter, Emulsifier (Soya Lecithins)], Sugar, Fortified Wheat Flour [Wheat Flour, Calcium Carbonate, Iron, Niacin (B3), Thiamin (B1)], ***Egg***, Dextrose, Glucose Syrup, Concentrated Orange Juice (1%), Gelling Agents (Citric Acid, Pectins), Humectant (Glycerol), Rapeseed Oil, Raising Agents (Ammonium Carbonates, Diphosphates, Sodium Carbonates), Acidity Regulators (Sodium Citrates, Citric Acid), Flavourings, Colour (Curcumin), Wheat Bran	Glucose-Fructose Syrup, Dark Chocolate (22%)(Sugar, Cocoa Mass, Cocoa Butter, Emulsifier (Soya Lecithins)), Sugar, Wheat Flour (Wheat Flour, Calcium Carbonate, Iron, Niacin, Thiamin), ***Pasteurised Egg***, Dextrose, Glucose Syrup, Concentrated Orange Juice (1%), Gelling Agents (Pectin, Citric Acid), Humectant (Glycerine), Sunflower Oil, Raising Agents (Ammonium Bicarbonate, Disodium Diphosphate, Sodium Bicarbonate), Acidity Regulators (Sodium Citrate, Citric Acid), Flavouring, Colour (Curcumin), Wheat Bran.

✗ indicates that an ingredient was not added to the corresponding sample. Other ingredients that may have influenced differences in product attributes across the samples are written in bold.

B.1.1 Sample Appearance and Presentation

Figure B 1. Photo of Jaffa cake samples showing variation in appearance.
Brand A exhibits greater variation in shape compared to Brands B and C. Samples were labelled with 3-digit codes during the study.

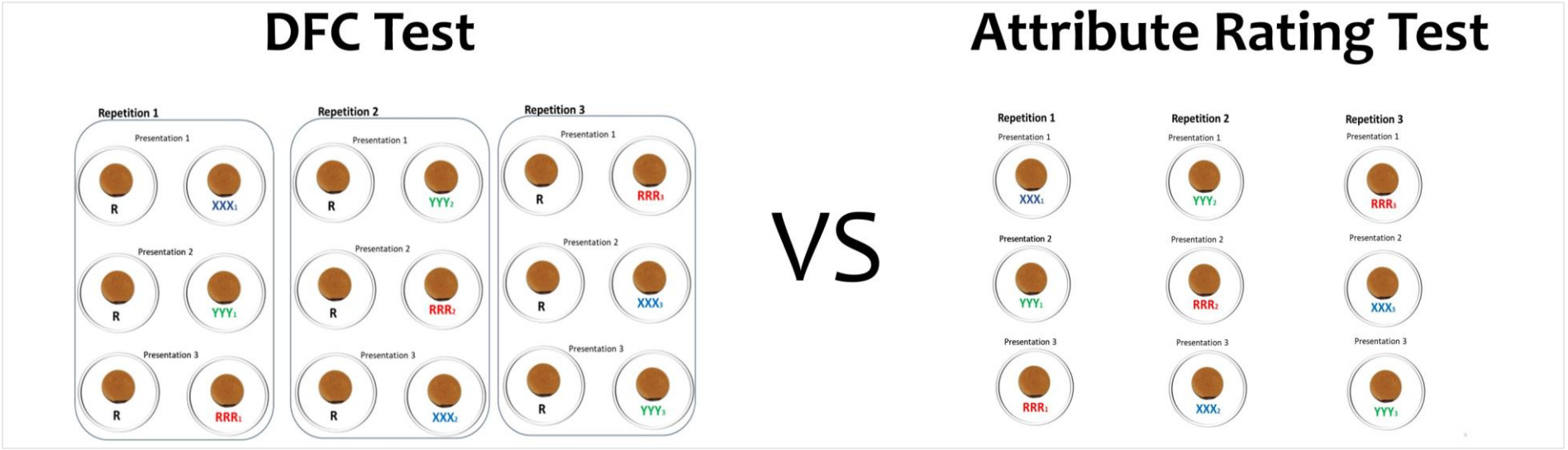


Figure B 2. Illustration of the presentation of Jaffa cake samples in the Difference from Control (DFC) and Attribute Rating (AR) tests.

B.2 Chocolate Spread Sample Content

Table B 2. Sample composition for the three chocolate spread samples in the Chapter 5 study

Attributes of Interest	***Content per 100g***	**Brand A**	**Brand B**	**Brand C**
Orange Flavour ***Sweetness*** ***Cocoa Flavour*** ***Milky Flavour*** ***Saltiness***	Energy (Kcal)	568	572	492
	Total Fat (g)	35.8	38	39
	Saturates(g)	14.5	8.7	9.1
	Total Carbs (g)	56.6	54	50
	Sugars (g)	54.7	52	<0.5
	Fibre (g)	2	1.7	4.4
	Protein (g)	3.6	3.8	3.1
	Salt (g)	0.13	0.1	<0.01
Comparison of Orange, Cocoa, and Milk flavour Content	**ORANGE FLAVOURING**			
	Orange Extract	✓	✗	✗
	Natural Orange Flavouring	✗	✗	✓
	COCOA CONTENT			
	Fat Reduced Cocoa Mass	12%	5%	14%
	Milk Chocolate Crumb	✗	7%	✗
	MILK CONTENT			
	Whey Powder	5%	✓	✗
	Lactose	✓	✓	✗
	Full Cream Milk Powder	✗	7%	✗
List of Ingredients		***Sugar***, Palm Oil, Fat Reduced Cocoa Powder (12%), Whey Powder (Milk) (5%), Lactose (Milk), Hazelnuts, Emulsifier (Sunflower Lecithin), Orange Extract, Flavouring	***Sugar***, Rapeseed Oil, Lactose (Milk), Palm Oil, Milk Chocolate Crumb (7%) (Milk, Sugar, Cocoa Mass), Full Cream Milk Powder (7%), Fat Reduced Cocoa (5%), Whey Powder (Milk), Emulsifier: Sunflower Lecithin, Flavouring	***Natural Sweetener (Maltitol)***, Vegetable Oil (Rapeseed, Sustainable Palm), Fat Reduced Cocoa 14%, Emulsifier (Sunflower Lecithin), Natural Orange Flavouring

✓ indicates that an ingredient is present in the sample, but its quantity was not specified on the label, while ✗ indicates that the ingredient was not added to the corresponding sample. Other ingredients that may have influenced differences in product attributes across the samples are written in bold.

B.3 Tomato Soup Sample Content

Table B 3 Sample composition for the three tomato soup samples in the Chapter 6 study

Sample Content		Sample A	Sample B	Sample C/Control
	BASE SOUP	**Cream of Tomato**	**Cream of Tomato and Basil**	**Cream of Tomato and Basil**
Extra ingredients added to base soup	**Dried Chopped Basil**	9%	✗	✗
	Tomato passata	✗	3.8%	✗
	Double cream	2%	✗	✗
	Garlic granules	✗	6.3%	✗
List of ingredients in base soup		**Tomatoes (89%)**, Water, Modified Cornflour, Sugar, Rapeseed Oil, Dried Skimmed Milk, Salt, Cream (Milk), Milk Proteins, Acidity Regulator - Citric Acid, Spice Extracts, Herb Extract	**Tomatoes (84%)**, Water, **Basil**, Herbs, Modified Cornflour, Sugar, Rapeseed Oil, Dried Skimmed Milk, Salt, Cream (Milk), Milk Proteins, Acidity Regulator - Citric Acid, Spice Extracts, Herb Extract	**Tomatoes (84%)**, Water, **Basil**, Herbs, Modified Cornflour, Sugar, Rapeseed Oil, Dried Skimmed Milk, Salt, Cream (Milk), Milk Proteins, Acidity Regulator - Citric Acid, Spice Extracts, Herb Extract
Photos showing side and top views of ready-to-serve samples				
Attributes of Interest				
Appearance	Glossy appearance, Viscous appearance, Colour intensity, Herby appearance			
Aroma	Rich Aroma, Cooked tomato aroma, Pungent aroma			
Mouthfeel	Smooth mouthfeel, Homogeneous mouthfeel, Thick mouthfeel			
Flavour	Savoury flavour, Herbal flavour, Cooked tomato flavour, Creamy flavour			
Taste	Salty taste, Sour taste, Sweet taste, Aftertaste			

✗ indicates that an ingredient was not added to the corresponding sample. Other ingredients that may have influenced differences in product attributes across the samples are written in bold.

Appendix C : Sensory Study Questionnaires

C.1 Jaffa cakes study questionnaire (RedJade)

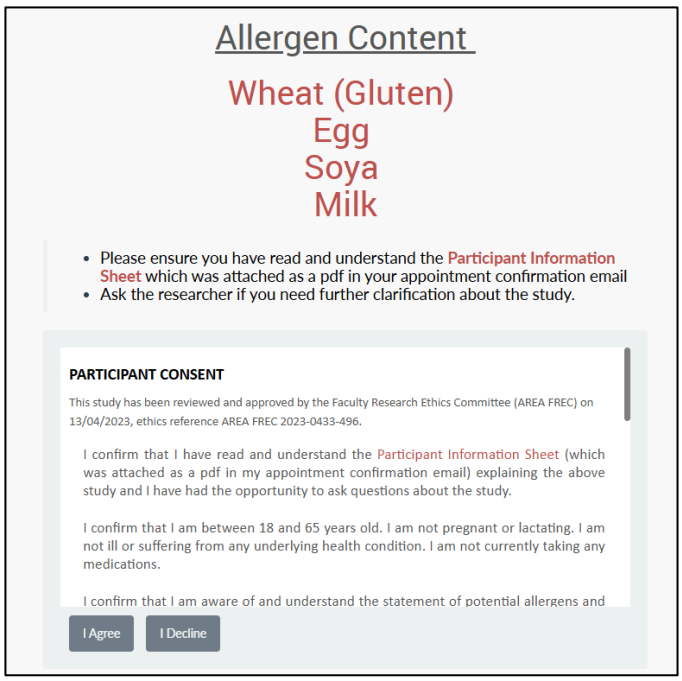


Figure C 1. Questionnaire introductory page for Jaffa cakes study. Participants must scroll through the entire consent form before they are able to select the "I Agree, or I Decline" button.

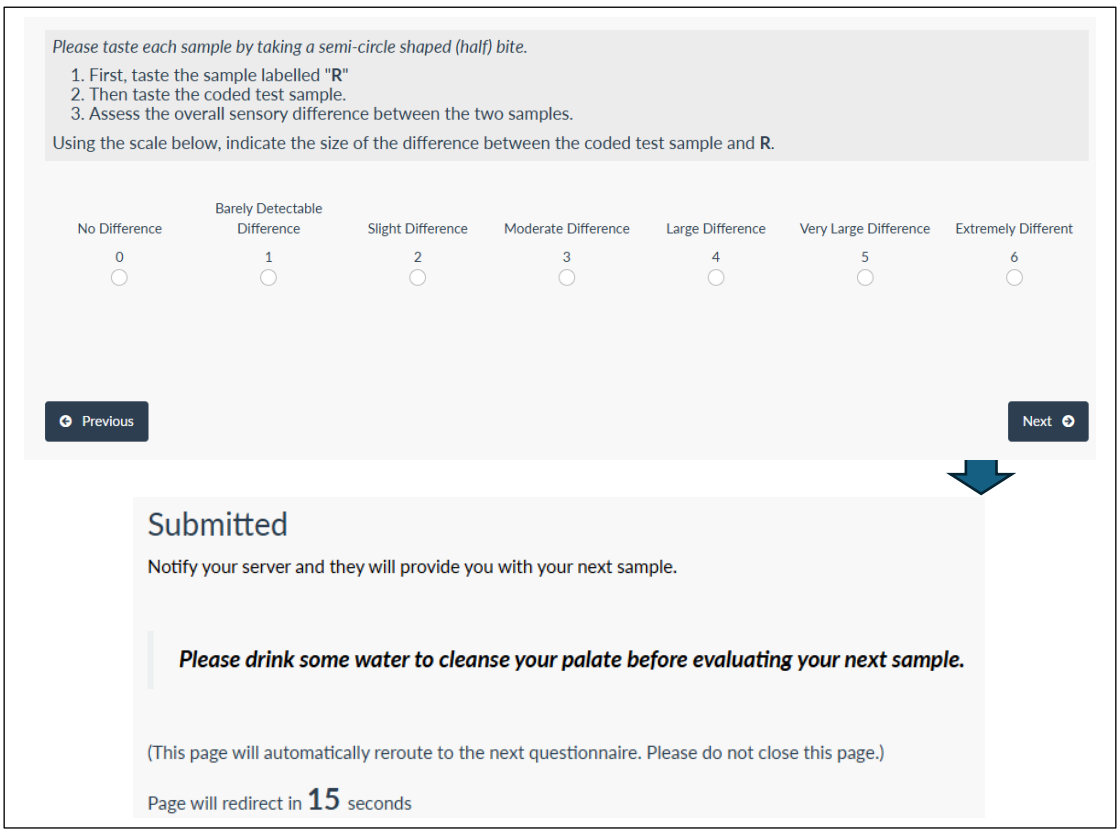


Figure C 2. Screenshot of DFC test questionnaire for Jaffa cakes study. The blue arrow signifies the transition to the next page.

Figure C 3. Screenshot of attribute rating questionnaire for Jaffa cakes study. The order of questions was randomised for each sample across all assessors.

C.2 Chocolate spread study questionnaires (RedJade)

Welcome

The following is the Study Participant Agreement. Please indicate your acceptance of the terms and conditions of the agreement by clicking the "I Agree" button at the bottom of the agreement.

STUDY PARTICIPANT AGREEMENT

You are being invited to participate in a research study titled '*Application of the Rasch Model on Sensory Panels*'. This study is being done by *Victoria Gill* from the University of Leeds.

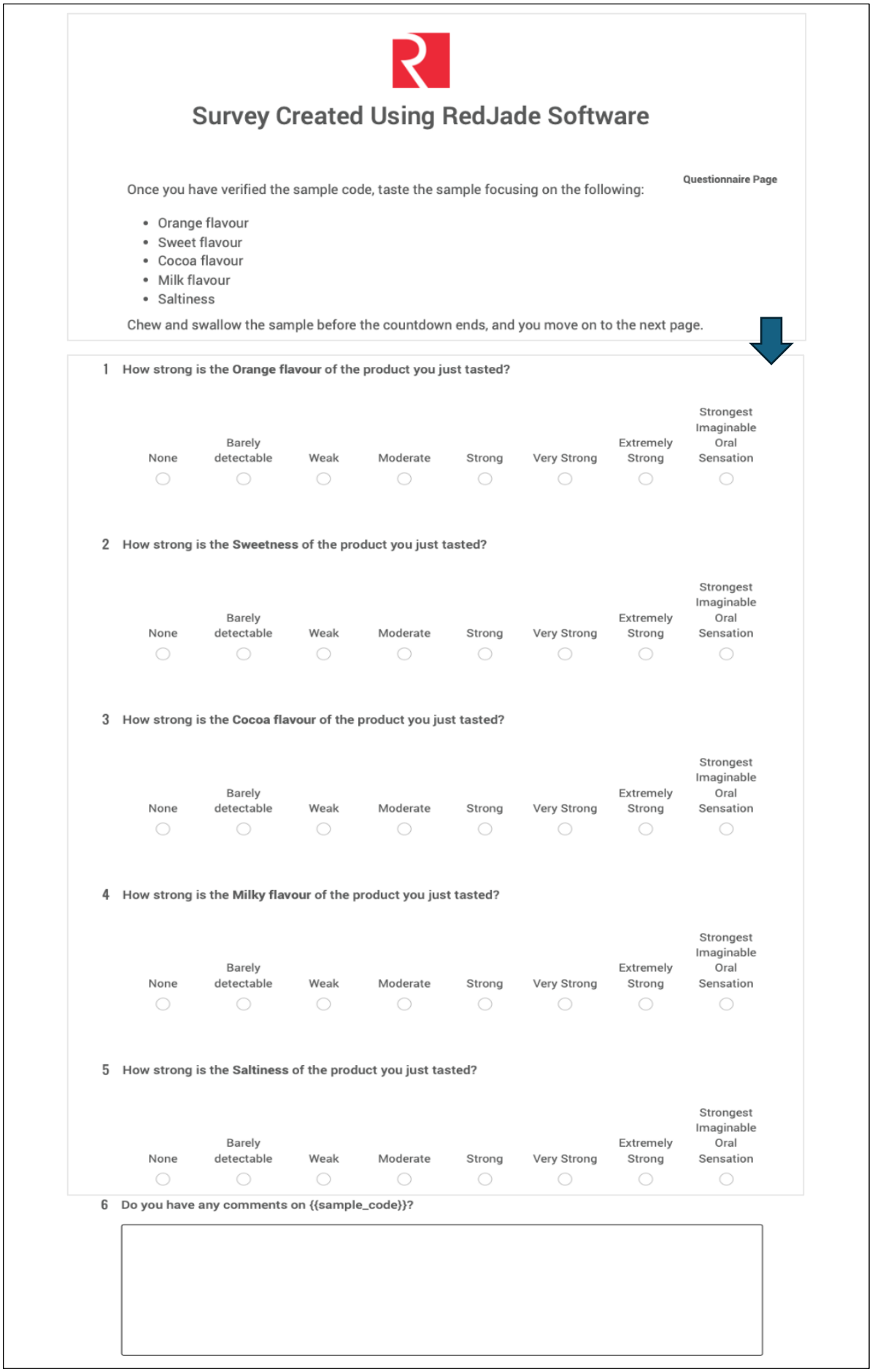
The purpose of this research study is *to determine how the Rasch model can be applied to different sensory panels to improve panellist performance* and will take you approximately *6 hours* to complete. Your participation in this study is entirely voluntary and *can be withdrawn at any time before the 31st December 2022 by emailing ll18vcg@leeds.ac.uk*. You do not have to answer any questions you do not want to.

We believe there are no known risks associated with this research study; however, as with any online-related activity, the risk of a breach is always possible. To the best of our ability, your participation in this study will remain confidential, and only anonymised data will be published. We will minimise any risks by *using RedJade software to anonymise results as they are created*. Further information is available via the University of Leeds.

By clicking the "I Agree" button below, I accept the terms and conditions of this agreement and confirm I agree with the following points:

- I confirm that I have read and understood the above information explaining the research project, and I have had the opportunity to as questions about the project
- I understand that my participation is voluntary and that I am free to withdraw without giving any reason until **31st December 2022** and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline. Please contact Victoria Gill at ll18vcg@leeds.ac.uk should you want to withdraw.
- I understand that members of the research team may have access to my **anonymised** responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.
- I understand that my responses will be kept strictly confidential
- I understand that the data collected from me may be stored and used in relevant future research **in an anonymised form**
- I understand that relevant sections of the data collected during the study may be looked at by individuals from the University of Leeds or from regulatory authorities where it is relevant to my taking part in this research.
- I agree to participate in the above research project

Figure C 4. Questionnaire introductory page and participant consent form for chocolate spread study.



Survey Created Using RedJade Software

Questionnaire Page

Once you have verified the sample code, taste the sample focusing on the following:

- Orange flavour
- Sweet flavour
- Cocoa flavour
- Milk flavour
- Saltiness

Chew and swallow the sample before the countdown ends, and you move on to the next page.

1 How strong is the **Orange flavour** of the product you just tasted?

None	Barely detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable Oral Sensation
○	○	○	○	○	○	○	○

2 How strong is the **Sweetness** of the product you just tasted?

None	Barely detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable Oral Sensation
○	○	○	○	○	○	○	○

3 How strong is the **Cocoa flavour** of the product you just tasted?

None	Barely detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable Oral Sensation
○	○	○	○	○	○	○	○

4 How strong is the **Milky flavour** of the product you just tasted?

None	Barely detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable Oral Sensation
○	○	○	○	○	○	○	○

5 How strong is the **Saltiness** of the product you just tasted?

None	Barely detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable Oral Sensation
○	○	○	○	○	○	○	○

6 Do you have any comments on {{sample_code}}?

Figure C 5. Screenshot of attribute rating questionnaire for chocolate spread study. The blue arrow signifies the transition to the next page.

C.3 Preview of test procedure for tomato soup sensory study

Your 1st Test is the Attributes Intensity Rating Test

At your test session, you will receive:

- **A unique assessor code:** This code will be used throughout the study.
- **A paper copy of the consent form:** You will sign two copies, one for you to keep and one for the researcher, who will also countersign it.
- **A test instruction document:** Please read this document before we start the test

What to expect

Tasting Periods: There will be 3 tasting periods in this test, each assigned to a different tab on the web browser.

- You will have a 5-minute break between each period.
- After completing a tasting period, the next one will be in the next tab in your browser.

Servings: During each tasting period, you will receive 3 test samples presented one at a time.

Sample Evaluation: Your task will be to evaluate and rate the intensity of several sensory attributes for each sample in the following order:

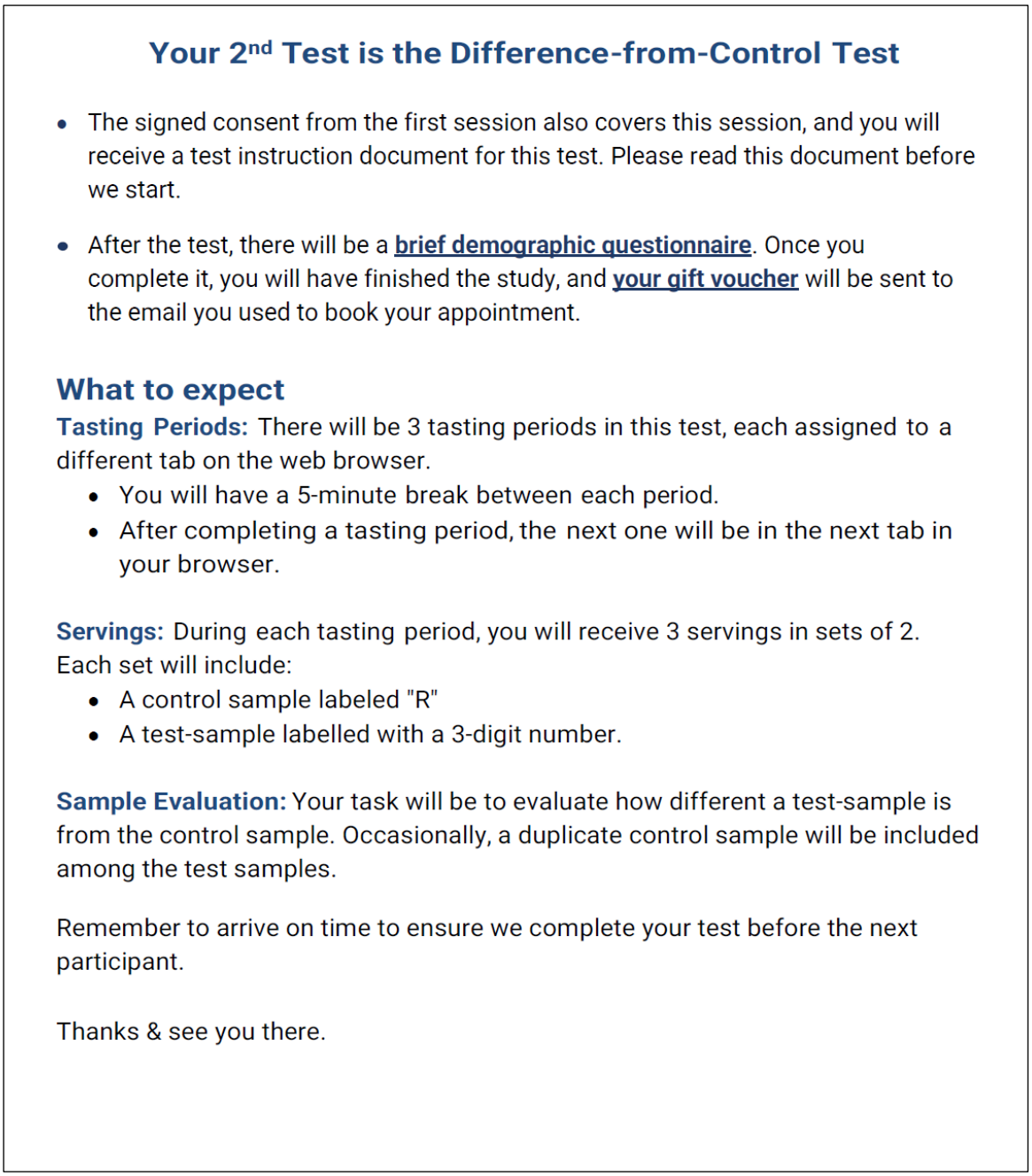
1. **Look:** examine the sample at eye level and rate the intensity of the appearance attributes.
2. **Smell:** Next, smell it and rate the intensity of the aroma attributes.
3. **Taste:** Finally, take three sips of the sample, one for each of the following steps:
    - *First sip: Assess and rate what it feels like in the mouth (mouthfeel/consistency/texture attributes).*
    - *Second sip: Assess and rate what flavours you perceive (flavour attributes).*
    - *Third sip: Assess and rate what it tastes like (taste attributes).*

Sensory Attributes: A written description of the specific attributes to look for will be provided during the test.

Remember to arrive on time to ensure we complete your test before the next participant.

Thanks & see you there.

Figure C 6. Screenshot of the preview document for the AR test



Your 2nd Test is the Difference-from-Control Test

- The signed consent from the first session also covers this session, and you will receive a test instruction document for this test. Please read this document before we start.
- After the test, there will be a **brief demographic questionnaire**. Once you complete it, you will have finished the study, and **your gift voucher** will be sent to the email you used to book your appointment.

What to expect

Tasting Periods: There will be 3 tasting periods in this test, each assigned to a different tab on the web browser.

- You will have a 5-minute break between each period.
- After completing a tasting period, the next one will be in the next tab in your browser.

Servings: During each tasting period, you will receive 3 servings in sets of 2. Each set will include:

- A control sample labeled "R"
- A test-sample labelled with a 3-digit number.

Sample Evaluation: Your task will be to evaluate how different a test-sample is from the control sample. Occasionally, a duplicate control sample will be included among the test samples.

Remember to arrive on time to ensure we complete your test before the next participant.

Thanks & see you there.

Figure C 7. Screenshot of the preview document for the DFC test

C.4 Tomato soup study questionnaires (RedJade)

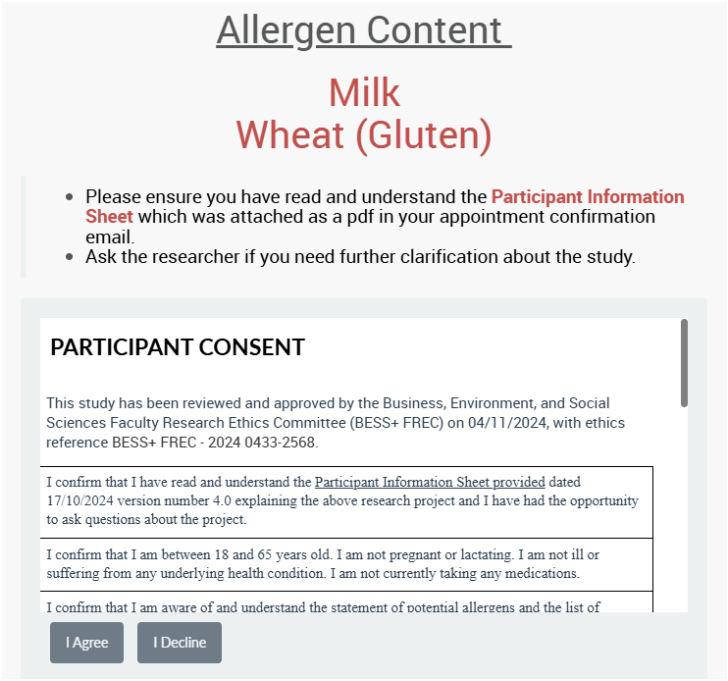


Figure C 8. Questionnaire introductory page for tomato soup study. Participants must scroll through the entire consent form before they are able to select the "I Agree, or I Decline" button.

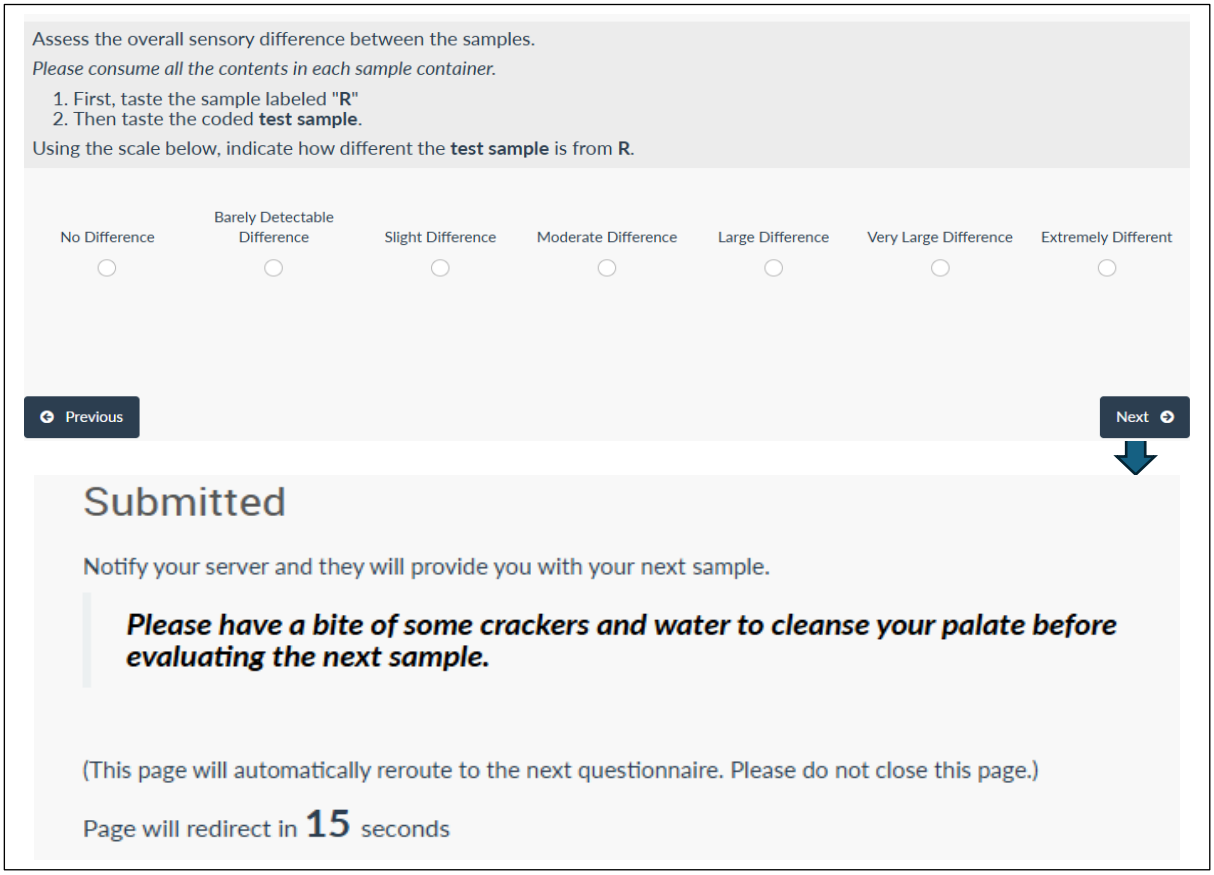


Figure C 9. Screenshot of DFC test questionnaire for tomato soup study. The next page after each questionnaire shows the palate cleanse instruction with a mandatory 15 secs timer before the next question.

Thank You!
Please think about the products you have evaluated and on the next page, identify the reasons for any differences you detected.

Previous | Next

Each time, you were presented with two sets of samples to identify differences from the control sample, R.
Based on the list below, did you notice any differences between the **Test Samples** and **R**?
For example, if you thought the glossy appearance of any of the test samples was different from that of R, you would select "YES" for glossy appearance. If you thought it was not different, you would select "NO".

	NO	YES
Glossy Appearance (degree of shine or reflected light from the surface)	○	○
Herby Appearance (presence of small, chopped pieces of herbs)	○	○
Colour Intensity (intensity or strength of colour from light to dark)	○	○
Viscous Appearance (thick and slow-moving)	○	○
Pungent Aroma (sharp, physically penetrating sensation in the nasal cavity)	○	○
Cooked Tomato aroma (typical smell of cooked tomato)	○	○
Rich Aroma (combination of multiple ingredients creating a deep and full aroma e.g. well-seasoned food)	○	○
Creamy Flavour (flavour associated with dairy products e.g cream, cheese)	○	○
Savoury Flavour (rich, spicy flavour associated with vegetable stock or meat broth)	○	○
Herbal Flavour (underlying flavour of dried herbs e.g. basil, oregano)	○	○
Smooth Mouthfeel (felt velvety or silky in the mouth, not rough or grainy)	○	○
Homogenous Mouthfeel (felt the same way throughout)	○	○
Thick Mouthfeel (felt dense or heavy in the mouth)	○	○
Cooked Tomato Flavour (typical cooked tomato flavour)	○	○
Sweet Taste (typical sweet taste e.g sugar/sucrose)	○	○
Sour Taste (sharp, tangy or tart taste e.g. citric acid in lemons)	○	○
Salty Taste (typical salt flavour e.g. common salt/NaCl or seawater)	○	○
Aftertaste (residual taste in the mouth after ingestion)	○	○

Please comment on any other differences you noticed that are not on the list.

Figure C 10. Final stage of the DFC test questionnaire for tomato soup, completed after all repeated sessions, requesting additional information on perceived attribute differences between samples.

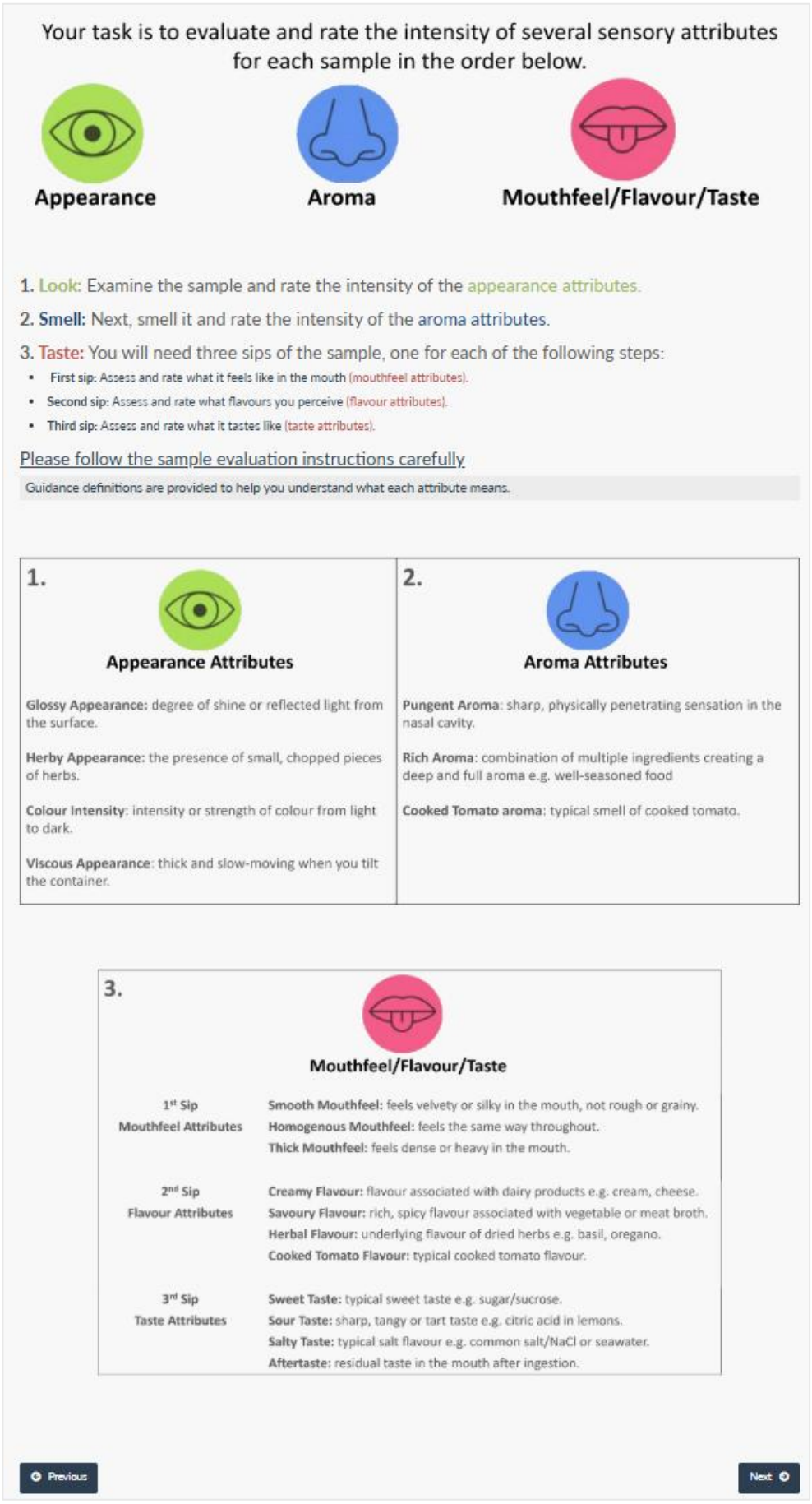


Figure C 11. Instruction page for the tomato soup attribute rating test, presented at the beginning of the test prior to sample presentation.

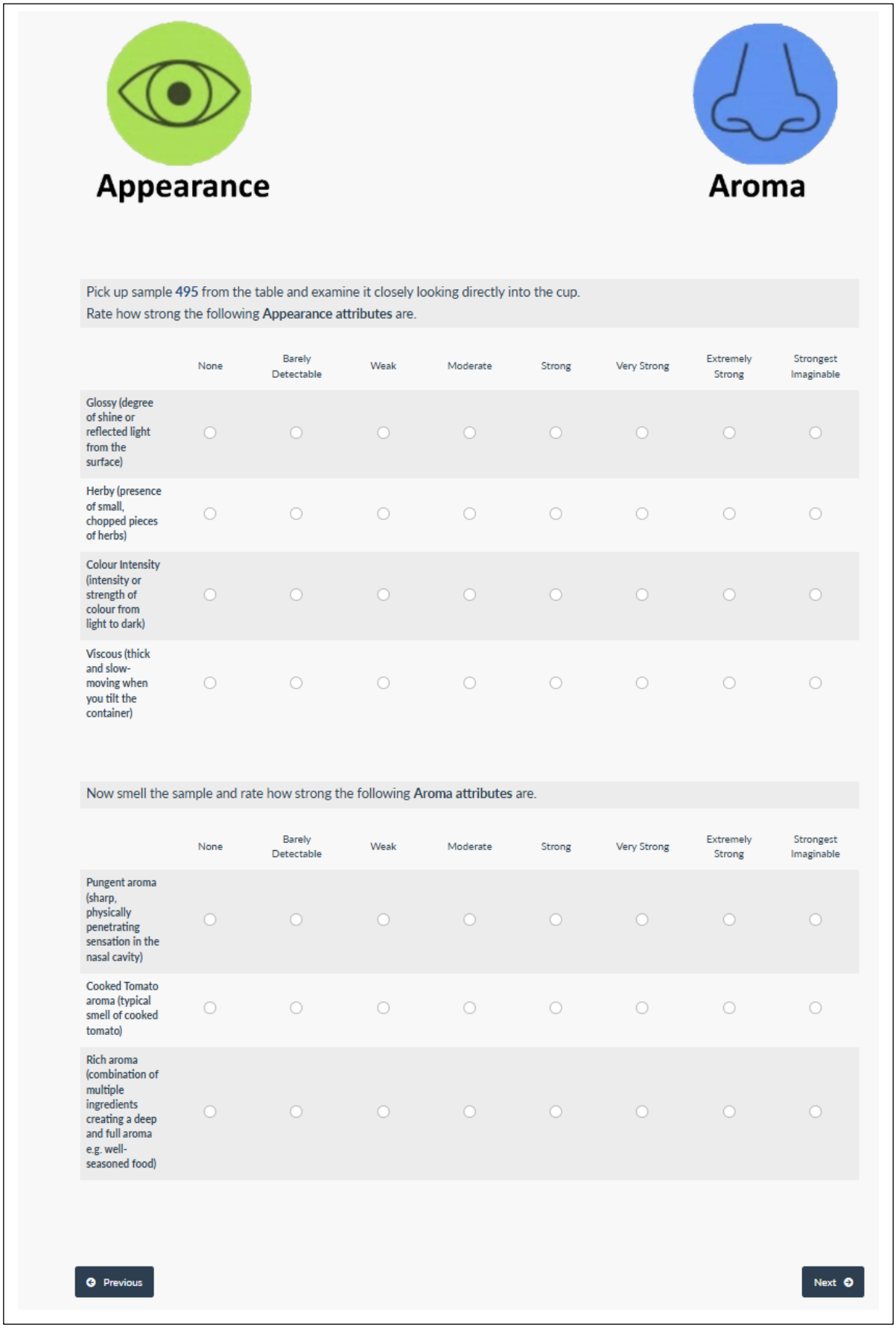
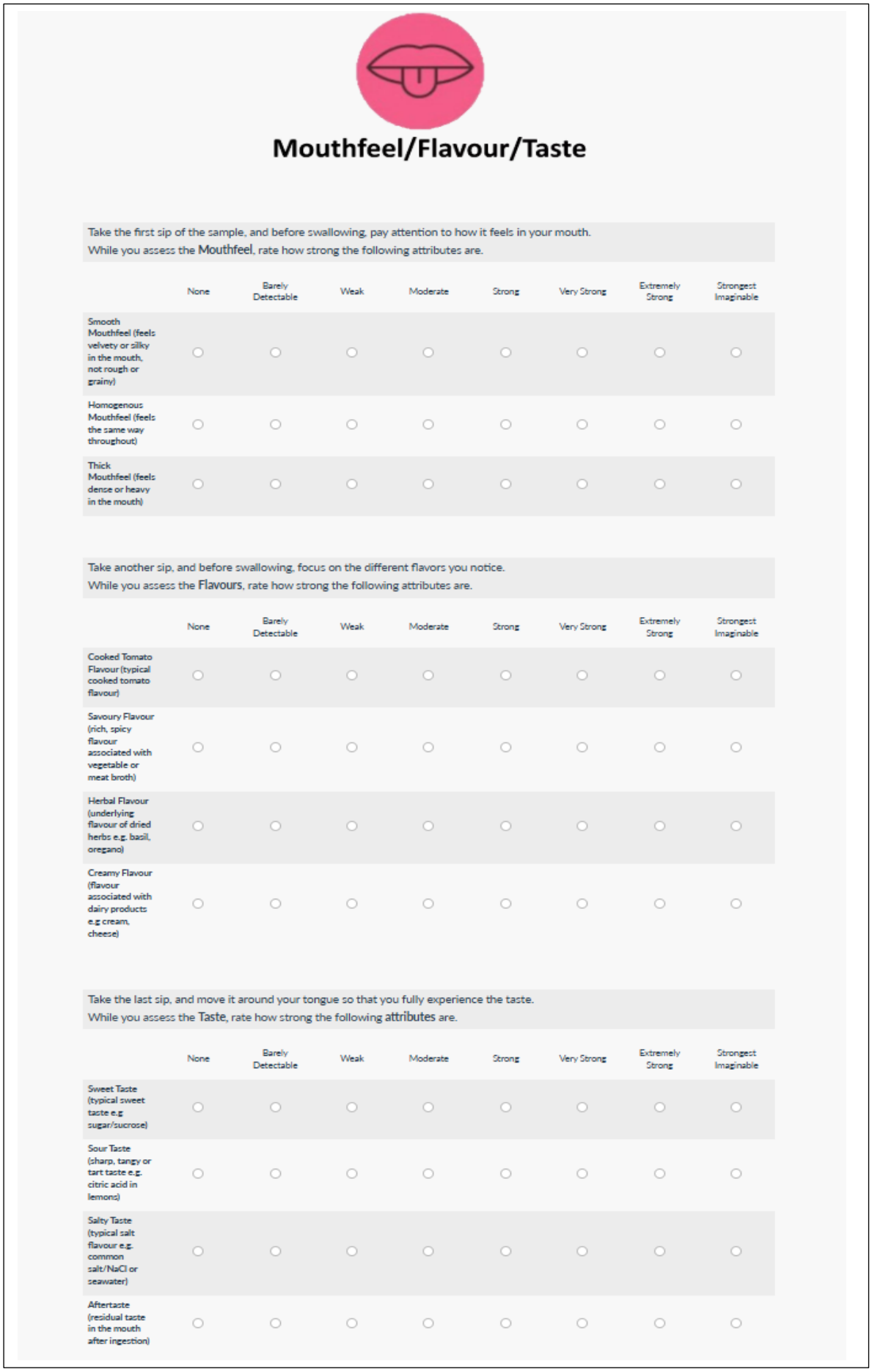


Figure C 12. Page 1 of attribute rating test questionnaire for tomato soup showing only appearance and aroma attributes.



Mouthfeel/Flavour/Taste

Take the first sip of the sample, and before swallowing, pay attention to how it feels in your mouth.
While you assess the **Mouthfeel**, rate how strong the following attributes are.

	None	Barely Detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable
Smooth Mouthfeel (feels velvety or silky in the mouth, not rough or grainy)	○	○	○	○	○	○	○	○
Homogenous Mouthfeel (feels the same way throughout)	○	○	○	○	○	○	○	○
Thick Mouthfeel (feels dense or heavy in the mouth)	○	○	○	○	○	○	○	○

Take another sip, and before swallowing, focus on the different flavors you notice.
While you assess the **Flavours**, rate how strong the following attributes are.

	None	Barely Detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable
Cooked Tomato Flavour (typical cooked tomato flavour)	○	○	○	○	○	○	○	○
Savoury Flavour (rich, spicy flavour associated with vegetable or meat broth)	○	○	○	○	○	○	○	○
Herbal Flavour (underlying flavour of dried herbs e.g. basil, oregano)	○	○	○	○	○	○	○	○
Creamy Flavour (flavour associated with dairy products e.g cream, cheese)	○	○	○	○	○	○	○	○

Take the last sip, and move it around your tongue so that you fully experience the taste.
While you assess the **Taste**, rate how strong the following **attributes** are.

	None	Barely Detectable	Weak	Moderate	Strong	Very Strong	Extremely Strong	Strongest Imaginable
Sweet Taste (typical sweet taste e.g sugar/sucrose)	○	○	○	○	○	○	○	○
Sour Taste (sharp, tangy or tart taste e.g. citric acid in lemons)	○	○	○	○	○	○	○	○
Salty Taste (typical salt flavour e.g. common salt/NaCl or seawater)	○	○	○	○	○	○	○	○
Aftertaste (residual taste in the mouth after ingestion)	○	○	○	○	○	○	○	○

Figure C 13. Page 2 of attribute rating test questionnaire for tomato soup showing mouthfeel, flavour and taste attributes. The order of questions for the flavour attributes was randomised for each sample across all assessors.

Appendix D Rating Scale Category Statistics

Table D 1. Category statistics showing the use of the 8-category Intensity Scale by the panels in the chocolate spread study (Chapter 5)

Panel		Scale Categories	Frequency[1]	Average Measure[2]		OUTFIT Mnsq[3]	Rasch Andrich Threshold	
				Observed	Expected		Measure	Distance[4]
Trained	0	None	40 (13)	-2.06	-2.15	1.1		
	1	Barely detectable	23 (7)	-1.67	-1.54	0.8	-1.34	-0.60
	2	Weak	53 (17)	-0.79	-0.72	1.0	**-1.94°**	0.86
	3	Moderate	96 (30)	-0.34	-0.32	1.2	-1.08	1.00
	4	Strong	83 (26)	-0.03	-0.13	0.8	-0.08	1.49
	5	Very strong	19 (6)	**-0.11×**	0.01	1.1	1.41	1.61
	6	Extremely strong	**1 (1)***	**-0.22×**	0.15	1.1	3.02	
	7	Strongest imaginable oral sensation	**0 (0)***	—	—	—	—	
Untrained	0	None	104 (10)	-0.73	-0.73	1.0		
	1	Barely detectable	119 (11)	-0.63	-0.60	0.9	-0.80	**-0.09^**
	2	Weak	170 (16)	-0.48	-0.47	0.9	**-0.89°**	**0.19^**
	3	Moderate	228 (21)	-0.32	-0.34	1.3	-0.70	0.58
	4	Strong	194 (18)	-0.16	-0.22	0.8	-0.12	**0.19^**
	5	Very strong	153 (14)	-0.13	-0.11	1.0	0.07	**0.30^**
	6	Extremely strong	99 (9)	-0.06	-0.02	1.1	0.37	1.69
	7	Strongest imaginable oral sensation	13 (1)	**-0.07×**	-0.07	1.0	2.06	

[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 Mnsq 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 (i.e., the thresholds between adjacent scale categories. For 8 and 7 category scales, the minimum threshold distances are 0.51 and 0.57, respectively.

Unmet Criteria from 3.3.1.4: Rating scale category diagnostics

° Disordered category thresholds indicate that an adjacent category was never the most probable choice.

× Average measures do not advance along the latent variable.

* Less than 10 observations in category.

^ Minimum advancing distance <0.51.

...continued from Table D1

Panel		Scale Categories	Frequency	Average Measure		OUTFIT Mnsq	Rasch Andrich Threshold	
				Observed	Expected		Measure	Distance
Selected Untrained	0	None	38 (11)	-0.63	-0.73	1.1		
	1	Barely detectable	43 (12)	**-0.75**×	-0.6	0.6	-0.79	**0.22**^
	2	Weak	45 (13)	-0.39	-0.45	1.1	-0.57	**-0.36**^
	3	Moderate	78 (22)	-0.36	-0.31	1.5	**-0.93**°	0.94
	4	Strong	60 (17)	-0.14	-0.2	0.5	0.01	**-0.11**^
	5	Very strong	57 (16)	-0.09	-0.1	0.9	**-0.10**°	0.55
	6	Extremely strong	34 (9)	-0.05	-0.02	1.0	0.45	1.48
	7	Strongest imaginable oral sensation	**5 (1)***	**-0.06**×	0.04	1.2	1.93	

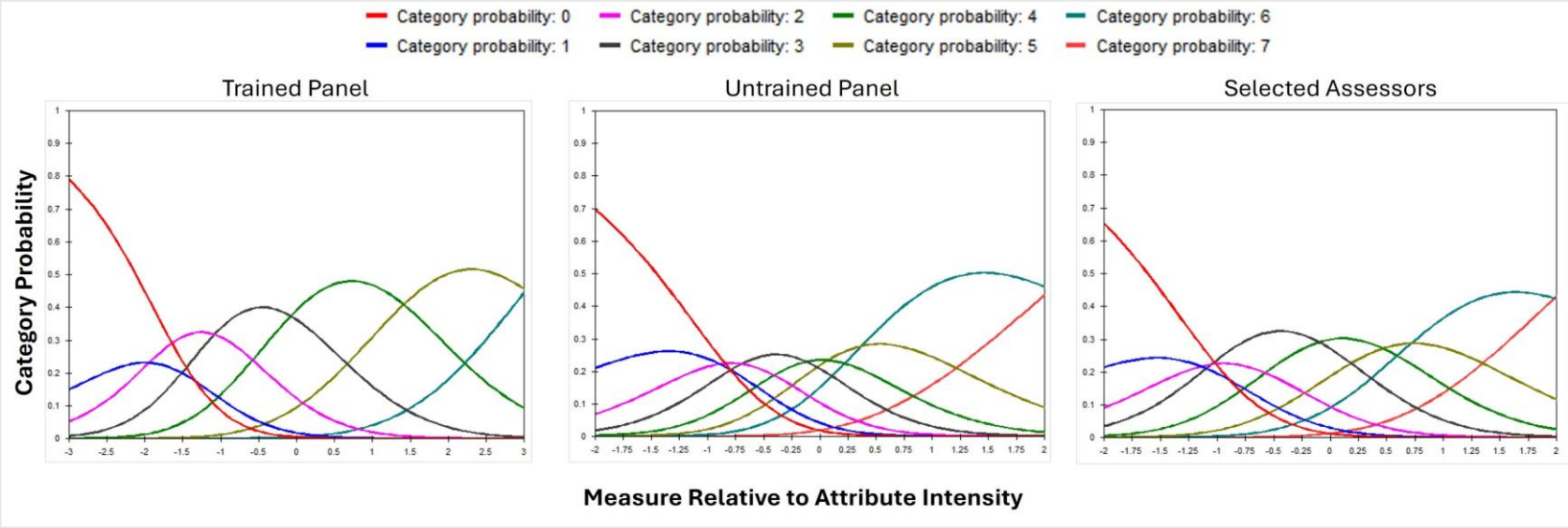


Figure D 1. Scale probability curves illustrating disordered Rasch-Andrich thresholds across all three panels.
The trained panel showed mild threshold disorder, primarily due to redundant scale categories that caused subtle undulation in the curves toward the upper end of the scale. In contrast, the full untrained panel and the selected assessors exhibited more disordered thresholds, indicating inconsistent use and poor distinction between the scale categories.

Appendix E Supplementary Statistics

Table E 1. Comparison of Rasch measures of product differences across all TIM-based datasets

	Model/Sample	Rasch χ^2	Parametric Test		Non-Parametric Test			
			Tukey's HSD×		Kruskal-Wallis		Friedman's	
			Rasch Mean		Mean Ranks		Rank Sum	
Jaffa Cakes, n =67 (Chapter 4)	**All Reps**	**17.9*****		**p=2.3e-07 *****		**p=7.5e-05*****		**p <2.2e-16*****
	Brand A		-0.59	a	1597.82	a	201	a
	Brand B		-0.71	ab~	1496.39	ab~	134	b
	Control		-0.78	b	1429.80	b	67	c
	Rep1	**5.4~**		**p=0.051~**		**p=0.139**		**p<2.2e-16*****
	Brand A		-0.67	a	526.97	a	201	a
	Brand B		-0.79	ab	499.05	a	134	b
	Control		-0.86	b	482.98	a	67	c
	Rep2	**4.6~**		**p=0.022***		**p=0.079~**		**p<2.2e-16*****
	Brand A		-0.59	a	528.57	a~	201	a
	Brand B		-0.68	ab	502.42	ab	134	b
	Control		-0.76	b	478.01	b~	67	c
	Rep3	**9.4****		**p=6.6e-05 *****		**p=0.003****		**p<2.2e-16*****
	Brand A		-0.67	a	544.44	a	201	a
	Brand B		-0.81	b	496.16	b	134	b
	Control		-0.89	b	468.40	b	67	c

Different superscript letters indicate significant difference (p<0.05), with ~ indicating marginal significance (p<0.1). P-value levels of significance: <0.001***, <0.01**, <0.05*, <0.1~; measures with no superscript symbols have p-values >0.1. All results from Rasch measures are based on the Kruskal-Wallis test. ***n***= total number of assessors in a panel.

× Sample differences were based on Tukey's HSD tests following a main effects ANOVA model: Sample + Assessor + Repetition.

...continued from Table E1

	Model/Sample	Rasch χ^2	Parametric Test		Non-Parametric Test		Model/Sample	
			Tukey's HSD×		Kruskal-Wallis		Friedman's	
			Rasch Mean		Mean Ranks		Rank Sum	
Choc. Spread (Chapter 5)	**Trained Panel (n=7)**	**7.5***		**p=0.008****		**p=5.3e-07*****		**9.1e-04*****
	Brand A		-0.51	a	180.55	a	21	a
	Brand B		-0.86	b	119.15	b	7	b
	Brand C		-0.54	a	174.30	a	14	c
	Untrained Panel (n=24)	**1.6**		**p=0.062~**		**p=0.113**		**p=3.8e-11*****
	Brand A		-0.30	a	568.02	a	72	a
	Brand B		-0.36	a	522.13	a	24	b
	Brand C		-0.35	a	531.35	a	48	c
Tomato Soup (Chapter 6)	**All Assessors (n=54)**	**106.3*****		**p< 2.2e-16*****		**p<0.001*****		**p< 2.2e-16*****
	Sample A		-0.07	a	5083.38	a	162	a
	Sample B		-0.21	b	4279.42	b	108	b
	Control		-0.30	c	3760.70	c	54	c
	Selection (n=17)	**27.5*****		**p< 2.2e-16*****		**p<0.001*****		**p=4.14e-08*****
	Sample A		-0.25	a	1557.18	a	51	a
	Sample B		-0.32	b	1424.17	b	34	b
	Control		-0.45	c	1151.15	c	17	c

Different superscript letters indicate significant difference ($p<0.05$), with ~ indicating marginal significance ($p<0.1$). P-value levels of significance: <0.001***, <0.01**, <0.05*, <0.1~; measures with no superscript symbols have p-values >0.1. All results from Rasch measures are based on the Kruskal-Wallis test. ***n***= total number of assessors in a panel.

× Sample differences were based on Tukey's HSD tests following a main effects ANOVA model: Sample + Assessor + Repetition.

Table E 2. Product mean comparisons across attributes for all TIM-based datasets

Jaffa Cakes (Chapter 4), (Nr =1005)

	Brand A	Brand B	Control	Attr. / Sa.[1]	*F. Av.*[2]	Measure
Orange Flavour	3.87[a]	3.42[b]	3.40[b]	**3.56**	***3.55***	0.75
Sweetness	3.71[a]	3.49[b]	3.46[b]	**3.55**	***3.54***	0.74
Cocoa Flavour	2.97[a]	3.11[a]	2.97[a]	**3.02**	***3.02***	0.21
Milky Flavour	2.30[a]	2.32[a]	2.26[a]	**2.29**	***2.30***	-0.50
Saltiness	1.71[a]	1.63[ab]	1.51[b]	**1.62**	***1.59***	-1.2
Sample / Attr. [3]	**2.91**	**2.80**	**2.72**			
Fair Average (F. Av.)	***2.90***	***2.79***	***2.72***			
Sample Logit	-0.59[a]	-0.71[ab]	-0.78[b]			

Chocolate Spreads (Chapter 5)

	Trained Panel (Nr = 105)						**Untrained Panel** (Nr =360)					
	Brand A	**Brand B**	**Brand C**	**Attr. / Sa.**	***F.Av.***	**Measr.**	**Brand A**	**Brand B**	**Brand C**	**Attr. / Sa.**	***F.AV.***	**Measr.**
Sweetness	3.76[a]	3.71[a]	3.48[a]	**3.65**	***3.65***	0.91	3.94[a]	5.04[b]	3.67[a]	**4.22**	***4.25***	0.47
Cocoa Flavour	3.43[b]	2.86[a]	3.29[ab]	**3.19**	***3.20***	0.39	3.85[b]	3.00[a]	3.92[b]	**3.59**	***3.60***	0.19
Orange Flavour	3.67[b]	1.67[a]	3.86[b]	**3.07**	***3.07***	0.26	2.62[a]	4.75[b]	2.42[a]	**3.26**	**3.27**	0.06
Milky Flavour	2.67[a]	3.38[b]	2.62[a]	**2.89**	***2.90***	0.10	3.64[b]	0.92[a]	3.60[b]	**2.72**	**2.71**	-0.16
Saltiness	0.57[a]	0.81[a]	0.71[a]	**0.70**	***0.68***	-1.66	1.94[a]	1.62[a]	1.86[a]	**1.81**	**1.78**	-0.55
Sample / Attr.	**2.82**	**2.49**	**2.79**				**3.20**	**3.07**	**3.09**			
Fair Average (F.Av)	***2.93***	***2.53***	***2.90***				***3.21***	***3.06***	***3.09***			
Sample Logit	-0.51[a]	-0.86[b]	-0.54[a]				-0.30[a]	-0.36[a]	-0.35[a]			

Values with different superscript letters indicate statistically significant differences ($p < 0.05$). **Nr** = Total number of responses used to estimate sample measures. **Sa. A** = Sample A; **Sa. B** = Sample B; **Ctrl** = Control. **Measr.** refers to the Rasch measure expressed on the logit scale, as shown in the Wright maps. Attributes are arranged in order from highest to lowest logit measure value.

[1] Attribute by Sample (Attr./ Sa): Raw mean scores of each attribute averaged across all samples.

[2] Fair average (F.Av): Rasch model expected score after adjusting for bias from other facets, and determine the relative position of the samples, or attributes on the logit scale.

[3] Sample by Attribute (Sample / Attr.): Raw mean scores of each sample averaged across all attributes.

Tomato Soups (Chapter 6)

	All Assessors (Nr = 2916)					
	Sa. A	**Sa. B**	**Ctrl**	***Attr./Sa.*[1]**	***F.Av*[2]**	**Measr.**
Smooth Mouthfeel	3.90[a]	3.88[a]	4.29[b]	**4.02**	***4.03***	0.47
Homogenous Mouthfeel	3.85[a]	3.98[a]	4.11[a]	**3.98**	***3.99***	0.44
Cooked Tomato Flavour	3.93[a]	3.82[a]	3.88[a]	**3.88**	***3.88***	0.37
Thick Mouthfeel	4.48[a]	3.25[b]	3.28[b]	**3.68**	***3.68***	0.22
Glossy Appearance	3.62[a]	3.51[a]	3.52[a]	**3.55**	***3.55***	0.14
Colour Intensity	3.53[a]	3.64[a]	3.41[a]	**3.53**	***3.53***	0.12
Savoury Flavour	3.69[a]	3.63[a]	3.22[b]	**3.51**	***3.52***	0.11
Cooked Tomato Aroma	3.48[a]	3.43[a]	3.38[a]	**3.43**	***3.43***	0.05
Herbal flavour	3.55[a]	3.64[a]	3.08[b]	**3.42**	***3.43***	0.05
Viscous Appearance	4.34[a]	2.86[b]	3.05[b]	**3.42**	***3.42***	0.05
Rich Aroma	3.40[a]	3.62[a]	3.03[b]	**3.35**	***3.36***	0.00
Aftertaste	3.45[a]	3.22[ab]	3.01[b]	**3.23**	***3.23***	-0.08
Creamy Flavour	3.53[a]	2.83[b]	2.63[b]	**2.99**	***2.99***	-0.25
Herby Appearance	3.43[a]	2.77[b]	2.70[b]	**2.97**	***2.97***	-0.27
Salty Taste	2.89[a]	2.91[a]	2.77[a]	**2.86**	***2.86***	-0.34
Sour Taste	2.85[a]	2.85[a]	2.80[a]	**2.83**	***2.83***	-0.36
Pungent Aroma	2.65[b]	3.18[a]	2.67[b]	**2.83**	***2.83***	-0.36
Sweet Taste	2.83[a]	2.86[a]	2.75[a]	**2.81**	***2.81***	-0.37
Sample / Attr.[3]	**3.52**	**3.33**	**3.20**			
Fair Average (F.Av)	***3.53***	***3.33***	***3.20***			
Sample Logit	-0.07[a]	-0.21[b]	-0.30[c]			

	Selected Assessors (Nr = 918)					
	Sa. A	**Sa. B**	**Ctrl.**	***Attr./Sa.***	***F.Av***	**Measr.**
Smooth Mouthfeel	4.02[a]	4.17[a]	4.39[a]	**4.20**	***4.21***	0.63
Homogenous Mouthfeel	3.75[a]	3.90[a]	4.14[a]	**3.93**	***3.94***	0.42
Thick Mouthfeel	4.67[a]	3.43[b]	3.67[b]	**3.92**	***3.93***	0.41
Glossy Appearance	3.78[a]	3.71[a]	3.67[a]	**3.72**	***3.73***	0.26
Savoury Flavour	3.75[ab]	3.80[a]	3.23[b]	**3.59**	***3.60***	0.17
Cooked Tomato Flavour	3.55[a]	3.67[a]	3.55[a]	**3.59**	***3.60***	0.16
Viscous Appearance	4.45[a]	3.00[b]	2.96[b]	**3.47**	***3.48***	0.08
Herbal flavour	3.39[ab]	3.80[a]	3.00[b]	**3.40**	***3.41***	0.03
Colour Intensity	3.20[b]	3.67[a]	3.30[ab]	**3.39**	***3.39***	0.02
Aftertaste	3.43[a]	3.35[a]	3.10[a]	**3.29**	***3.30***	-0.04
Rich Aroma	3.05[b]	3.67[a]	2.92[b]	**3.22**	***3.22***	-0.10
Creamy Flavour	3.67[a]	3.08[ab]	2.75[b]	**3.16**	***3.17***	-0.13
Cooked Tomato Aroma	3.12[a]	3.14[a]	3.16[a]	**3.14**	***3.14***	-0.15
Herby Appearance	3.55[a]	2.79[b]	2.75[b]	**3.03**	***3.03***	-0.23
Sour Taste	2.94[a]	2.82[a]	2.92[a]	**2.90**	***2.89***	-0.32
Pungent Aroma	2.57[b]	3.24[a]	2.63[b]	**2.81**	***2.81***	-0.38
Salty Taste	2.78[a]	2.82[a]	2.75[a]	**2.78**	***2.78***	-0.40
Sweet Taste	2.84[a]	2.78[a]	2.53[a]	**2.72**	***2.71***	-0.44
Sample / Attr.	**3.47**	**3.38**	**3.19**			
Fair Average (F.Av)	***3.49***	***3.40***	***3.20***			
Sample Logit	-0.25[a]	-0.31[b]	-0.45[c]			

Values with different superscript letters indicate statistically significant differences ($p < 0.05$). **Nr** = Total number of responses used to estimate sample measures. **Sa. A** = Sample A; **Sa. B** = Sample B; **Ctrl** = Control. **Measr.** refers to the Rasch measure expressed on the logit scale, as shown in the Wright maps. Attributes are arranged in order from highest to lowest logit measure value.

[1] Attribute by Sample (Attr./ Sa): Raw mean scores of each attribute averaged across all samples.

[2] Fair average (F.Av): Rasch model expected score after adjusting for bias from other facets, and determine the relative position of the samples, or attributes on the logit scale.

[3] Sample by Attribute (Sample / Attr.): Raw mean scores of each sample averaged across all attributes.

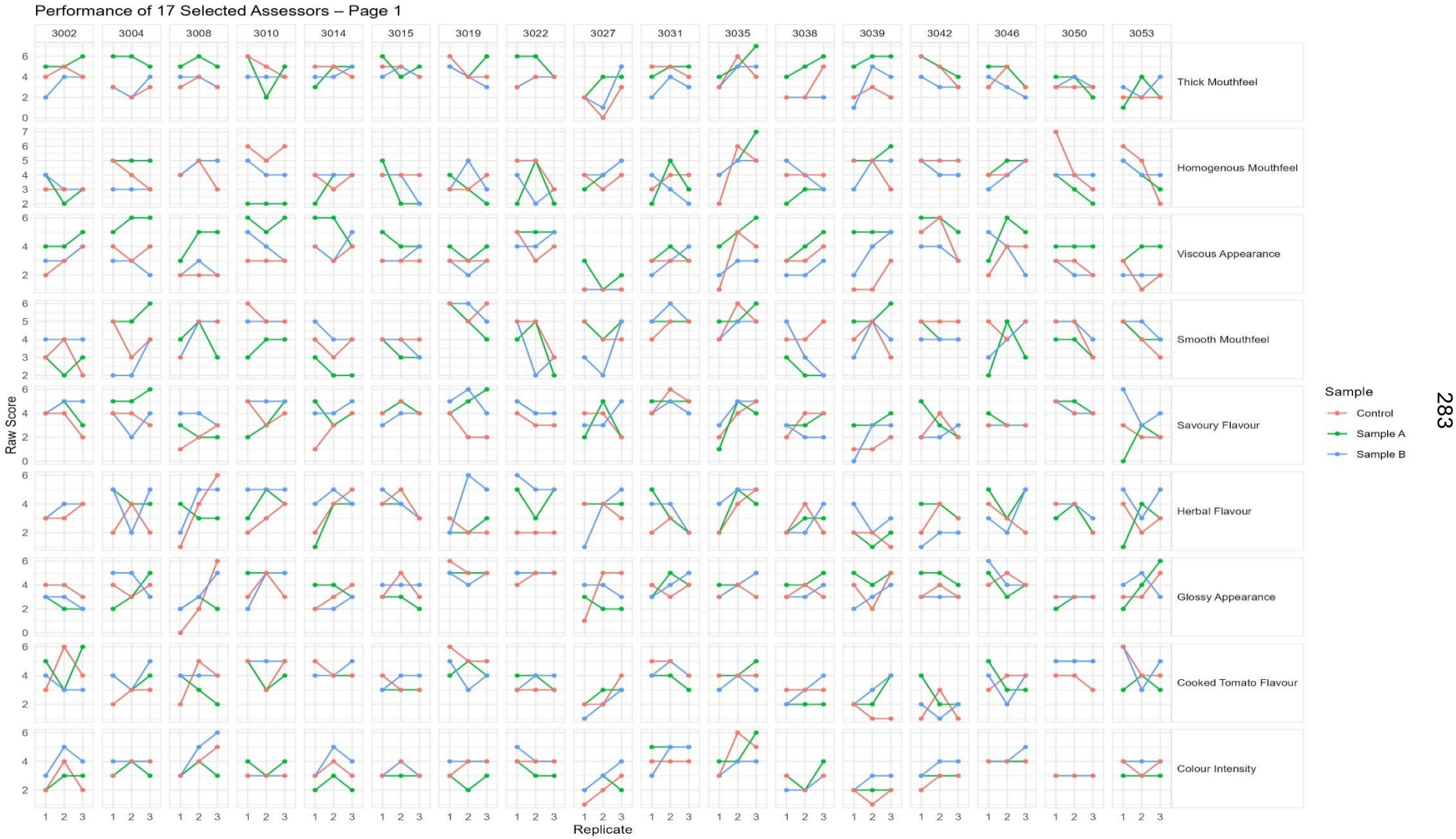


Figure E 1. Trellis plots showing individual response distributions for the subset of 17 assessors (Page 1 of 2). Attributes are arranged in the same order as presented in **Figure 5.5**

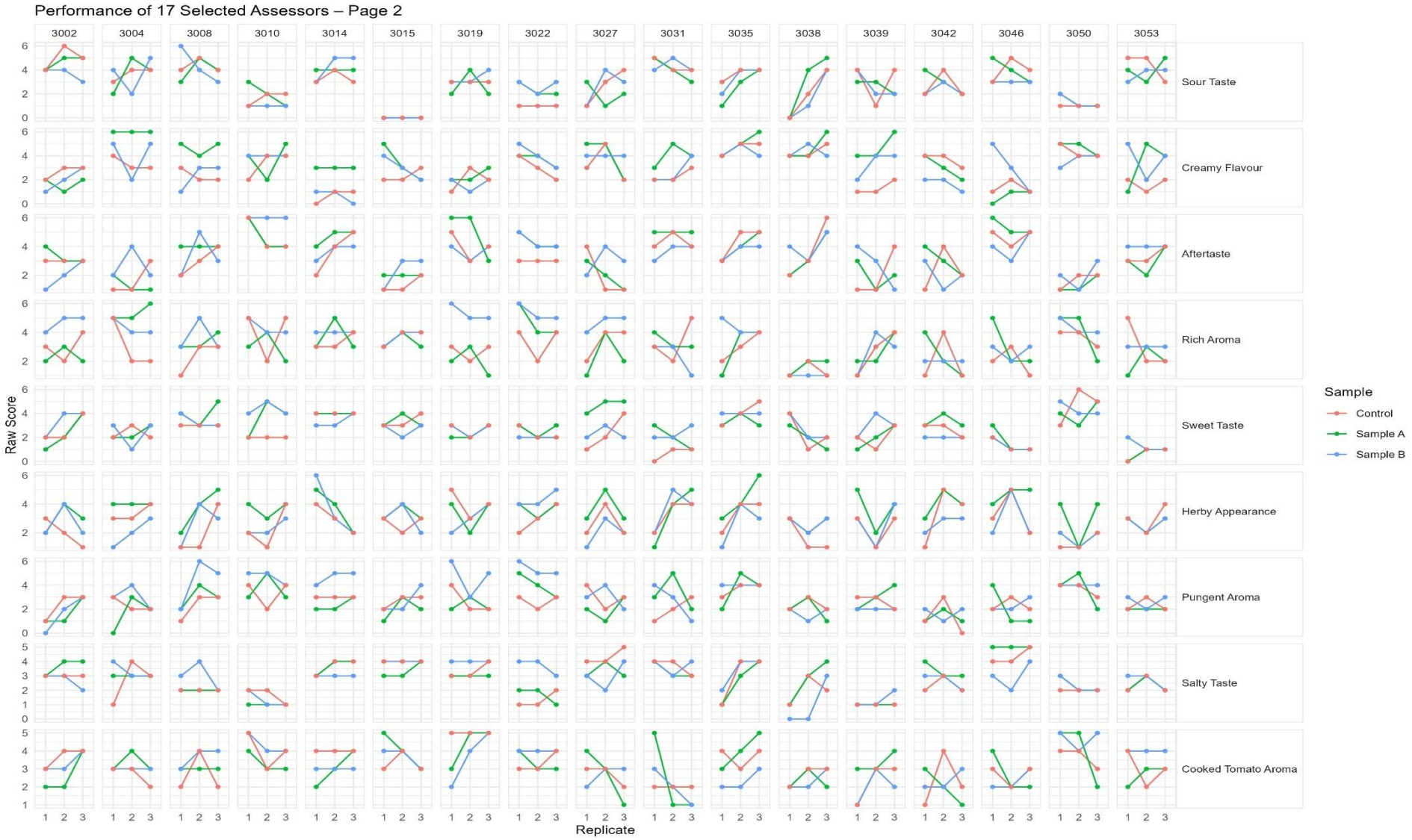


Figure E 1. Trellis plots showing individual response distributions for the subset of 17 assessors (Page 2 of 2). Attributes are arranged in the same order as presented in **Figure 5.5**

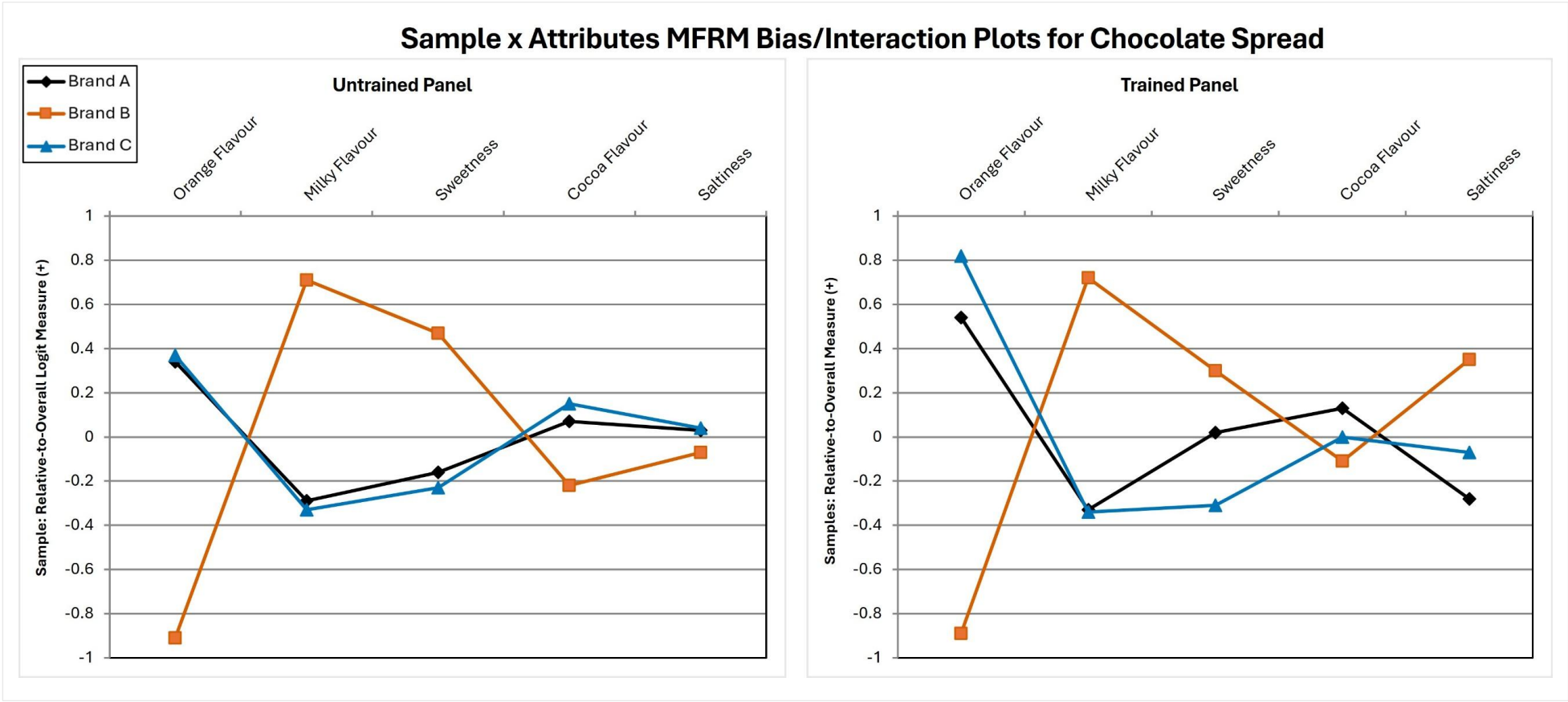


Figure E 2. Rasch-adjusted sample x attribute bias/interaction plots for the untrained panel (left) and trained panel (right).
The y-axis represents the relative-to-overall logit measure, calculated as the deviation of each sample's estimated logit rating on an attribute from that sample's overall logit measure in the Many-Facet Rasch Model. Positive values indicate attributes rated higher than expected based on the sample's overall measure, and negative values indicate attributes rated lower than expected. The plots reveal opposing attribute intensities, with Brand B displaying a strong negative deviation on Orange flavour and strong positive deviation on Milky flavour. These contrasts illustrate product-level differences that may be masked in the overall measure, particularly when inconsistent ratings dilute or cancel attribute-specific effects.