Article:
Ho, P (2019) A new approach to measuring Overall Liking with the Many-Facet Rasch Model. Food Quality and Preference, 74. pp. 100-111. ISSN: 0950-3293

https://doi.org/10.1016/j.foodqual.2019.01.015

A new approach to measuring Overall Liking with the Many-Facet Rasch Model

Peter Ho[a,*]

[a]*School of Food Science and Nutrition, University of Leeds, Leeds, LS2 9JT, United Kingdom*

Abstract

The 9-point hedonic scale is the most common hedonic rating scale used to provide an assessment of overall liking. Studies have shown that consumer judgements of overall liking could be influenced by their ratings of the liking of flavour, texture, aroma or appearance. However, this is not directly taken into account when using the holistic variable of overall liking. A new approach is proposed for measuring overall liking that is firstly based on initially considering what sensory characteristics (attributes or modalities) defines the latent sensory construct of OVERALL LIKING. The aim of this study was to develop a single measure of Overall Liking that incorporates the relative importance of liking ratings from different sensory characteristics by applying a Many-Facet Rasch model to produce interval-scaled estimates of Overall Liking. A homogeneity test found significant differences ($p<0.01$) between the Rasch means estimates of the cured 10 hams that were evaluated by a consumer panel ($n=90$), with the two different definitions of the Rasch measure of Overall Liking. No significant differences were found when comparing Rasch measures with raw scores using an intrablock BIB ANOVA and Durbin test. The *degree of Relevance*, shown on a Many-Facet Wright map, indicated the extent which a variable contributed to the measure of Overall Liking. Of the 10 sensory attributes used for the *Individual Attribute Measure*, Hardness and Juiciness contributed the most, while Sweetness and Typical Flavour contributed the least. However, the modalities used in rating the Likings of Overall Flavour, Texture, Aroma and Appearance contributed almost to the same extent to Overall Liking in the *Total Attribute Measure*. The Wright Map also showed that the categories on the 9-point hedonic scale were unequally spaced and the distance between them became increasingly larger the further away from the central cate-

*Corresponding author

Email address: p.ho@leeds.ac.uk (Peter Ho)

gory of ‘Neither like nor dislike’.

Keywords: 9-point hedonic scale, Overall Liking, Many-Facet Rasch Model, intrablock BIB ANOVA, Durbin test

1. Introduction

The 9-point hedonic scale has been widely used for many years to examine the overall liking of many different food products, such as coffee (Varela, Beltrán, & Fiszman, 2014), cookies (Mudgil, Baraka, & Khatkar, 2017; Rankin, Fada, & Bingham, 2000), dairy and non-dairy products (Bayarri, Carbonell, Barrios, & Costell, 2010; Drake & Gerard, 2003; Young, Drake, Lopetcharat, & McDaniel, 2004) and meat products (Pham et al., 2008; Shao, Avens, Schmidt, & Maga, 1999). However, a holistic rating of overall liking does not provide any indication of the contribution of sensory modalities or characteristics to overall liking. It is for this reason some studies examine the liking ratings of individual sensory characteristics to understand of the contribution they might have (Pham et al., 2008; Ramcharitar, Badrie, Marrfeldt-Beman, Matsuo, & Ridley, 2005; Young, Drake, Lopetcharat, & McDaniel, 2004). Often a trained sensory panel is used to provide a sensory profile and external preference mapping is used to relate the intensity of individual sensory attributes with the consumer panel’s ratings of overall liking (Hough & Sánchez, 1998; Pham et al., 2008; Young, Drake, Lopetcharat, & McDaniel, 2004). A few studies have used consumers, where regression methods have shown that individual sensory liking modalities (Liking of Appearance, Aroma, Flavour and texture) play an important contribution to the evaluation of overall liking (Andersen, Brockhoff, & Hyldig, 2019; Moskowitz & Krieger, 1995).

The original 9-point hedonic scale (Peryam & Girardot, 1952; Peryam & Pilgrim, 1957) has nine labelled categories, four distinct categories of dislike, a middle category ‘neither like nor dislike’ and four categories of liking. However, different variations of this scale have been used. The hedonic labels have been replaced by numerical values, word anchors used only at the extreme ends of the scale, or a combination of both words and numbers used on the same scale (Nicolas, Marquilly, & O’Mahony, 2010). Whatever choices made in those studies, this categorical scale produces, at best, ordinal-level data. The common practice of assigning numerical values from 1 to 9 to each category on this scale, from‘dislike extremely’ to ‘like extremely’, and analysing the collected raw data as interval-level data using parametric statistics is questionable. Parametric statistical anal-

ysis have found that data from the 9-point hedonic scale tends to violate certain statistical assumptions (Villanueva, Petenate, & Da Silva, 2000).

Researchers searching for ways to overcome some of these possible issues have developed a number of alternative category-ratio scales that produce interval-scaled data. The Labeled Affective Magnitude (LAM) scale is a continuous line scale with eleven unevenly spaced anchors, nine of them with labels corresponding to those used in the 9-point hedonic scale and two end anchors denoting 'Greatest Imaginable Dislike and 'Greatest Imagineable Like' (Schutz & Cardello, 2001). The LAM scale has been used in a number of studies to measure the overall liking of different foods (Bakke & Vickers, 2007; El Dine & Olabi, 2009; Forde & Delahunty, 2004). The LAM scale has been found to be comparable with the 9-point hedonic scale in measuring overall liking (Lawless, Cardello, et al., 2010; Lawless, Popper, & Kroll, 2010; Lawless, Sinopoli, & Chapman, 2010). The Labeled Hedonic Scale (LHS) and the hedonic general Labeled Magnitude Scale (gLMS) are two other category-ratio scales that have also been developed (Bartoshuk, Fast, & Snyder, 2005; Lim, Wood, & Green, 2009).

The main aim of this study was to examine the applicability of Many-Facet Rasch (MFR) models to sensory liking data, specifically, the construction of a single measure of Overall Liking. This paper examines how the MFR model, that is one of the key elements of the 'Four Building Blocks' construct modelling approach suggested by Wilson (2005), takes into account the contribution from the liking ratings of a set of individual attributes or sensory modalities in the estimation of Overall Liking.

It is proposed that a Rasch measure of Overall Liking would provide a more appropriate of measure of OVERALL LIKING than a single holistic rating of overall liking, consisting of either the liking ratings from a set of individual attributes or sensory modalities. Rasch modelling is also suggested as a preliminary step to parametric statistical analysis of data from a 9-point hedonic scale, producing interval-scaled data with properties similar to that of category-ratio scales, such as the LAM or LHS.

2. Defining the construct of OVERALL LIKING

A fundamental question that was considered in developing a Rasch measure of Overall Liking, was the question of 'How should this latent variable of Overall Liking be measured?'. A common approach is to use a single holistic variable called 'overall liking' to represent the construct OVERALL LIKING (Fig. 1), whereby an assessor rating a product sample might be asked 'How much do

you like product X?'. Liking responses from the original 9-point hedonic scale (Peryam & Girardot, 1952; Peryam & Pilgrim, 1957), from 'dislike extremely' to 'like extremely', would then be converted to raw scores, by assigning numerical values from 1 to 9. These raw overall liking scores would then be analysed by statistical methods and the estimates of mean values would represent the average degree of overall liking of the product being evaluated.

An alternative approach is to construct a single latent measure of Overall Liking, that is composed from the liking ratings of either a set of individual attributes or different sensory modalities, whereby each individual attribute or their sensory modality might differ in their *degree of Relevance* on the defined construct. i.e., the extent to which they contribute to the measure of Overall Liking. The contribution of liking ratings from individual attributes (Moskowitz & Krieger, 1993) or different sensory modalities (Moskowitz & Krieger, 1995) as drivers of overall liking have been previously modelled with multiple linear regression. Moskowitz and Krieger (1995) found liking of flavour as the most important contributor to overall liking, followed by liking of texture then liking of appearance for six different food categories. Similar results were found by Andersen, Brockhoff, and Hyldig (2019).

The construct modelling approach, suggested by Wilson (2005), has been used in the construction of the measure of Overall Liking. This approach differs from the common approach, previously mentioned, in a number of ways as shown in Fig. 2. In step (1), a construct map is created that is a visual representation of the unidimensional (single) construct, OVERALL LIKING. A list of ten individual attributes (Liking of Redness, Liking of Marbling, Liking of Typical Aroma, Liking of Hardness, Liking of Fibrousness, Liking of Juiciness, Liking of Saltiness, Liking of Sweetness, Liking of Typical Flavour, Liking of Aftertaste) or sensory modalities (Liking of Overall Appearance, Liking of Overall Aroma, Liking of Overall Texture, Liking of Overall Flavour) were chosen in step (2). These variables are possible components, that adequately represents the latent variable of OVERALL LIKING, used for the measure of Overall Liking. A set of questions were defined for each of them. For example, 'How much do you like the Hardness of product X?' or 'How much do you like the Overall Appearance of product X?'. Selecting the scale used for liking ratings was step (3). The 9-point hedonic scale with the categories, 'Dislike Extremely', 'Dislike Very Much', 'Dislike Moderately', 'Dislike Slightly', 'Neither Like or Dislike', 'Like Slightly', 'Like Moderately', 'Like Very Much', 'Like Extremely', were chosen to represent the *levels of Liking* on the Overall Liking measure. In step (4), a Many-Facet Rasch (MFR) model was chosen which modelled the probability of category responses,

instead of the raw scores used in the common approach for rating overall liking. Two measures of Overall Liking were evaluated in this study, an Individual Attribute Measure (IAM) composed of the ten individual attributes and a Total Attribute Measure (TAM) that was composed of the four sensory modalities. The Many-facet Rasch model described in eq. (4) was used to estimate these two measures. A third measure, the Single Overall Measure (SOM) which was composed of the single variable 'overall liking', served as a reference control. This measure was estimated using Eq. (5). A Wright Map was then used to relate measures of Overall Liking back to the construct OVERALL LIKING, by simultaneously representing all three 'facets' in a single plot. The Product facet indicates the *degree of Overall Liking* for each product from the least liked to most liked product. The Assessor facet measures the *degree of Criticality*, which represents the extent of liking judgements by an assessor, from less likely to dislike a product to more likely to dislike a product. Finally, the Criterion facet indicates the *degree of Relevance* or contribution the set of individual attributes or sensory modalities have to the measure, from more frequently chosen to less frequently chosen criterion. Rasch models are explained further in the next section.

3. Rasch measurement

Rasch models are latent trait measurement models for handling raw data scores from nominal and ordinal scales. Parameters are interval-scaled (Bond & Fox, 2015) on a logit-scale. i.e., the natural logarithm of the log-odds ratio, or simply 'log-odds units'. Nominal data can be modelled with the dichotomous Rasch Model, which was developed by the Danish mathematician George Rasch (Rasch, 1960). If a panel of sensory assessors (n) evaluating the overall liking of a single product in terms of a set of sensory attributes or sensory modalities (i), whereby each sensory attribute or sensory modality would be more frequently rated as liked ($x_{ni} = 1$) or disliked ($x_{ni} = 0$), then the log-odds form of the dichotomous Rasch Model can be represented by:

$$ln\left[\frac{P_{ni}}{1 - P_{ni}}\right] = \theta_n - \delta_i \tag{1}$$

where:

P_{ni} = probability that assessor (n) would rate a sensory attribute or modality (i) as liked
$1 - P_{ni}$ = probability that assessor (n) would rate a sensory attribute or modality (i) as disliked,
θ_n = *degree of Leniency* of assessor (n), where n=1,...,a,
δ_i = *degree of Relevance* for sensory attribute (i), where i=1,...,c.

The dichotomous Rasch Model is the simplest form of a family of unidimensional Rasch models, that includes two polytomous models. For a sensory test using a labelled category scale, like the 9-point hedonic scale, the Rating Scale (RS) Model (Andrich, 1978) can be used. A threshold parameter (τ_k) is included that models the position on the scale where there is a 50% probability that an assessor selects one of two adjacent scale categories (Eckes, 2011). The log-odds form of the RS model is

$$ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \delta_i - \tau_k \tag{2}$$

where:

P_{nik} = probability that assessor (n) rated k for sensory attribute or modality (i),
P_{nik-1} = probability that assessor (n) rated *k-1* for sensory attribute or modality (i),
θ_n = *degree of Leniency* of assessor (n), where n=1,...,a,
δ_i = *degree of Relevance* for sensory attribute (i), where i=1,...,c,
τ_k = boundary estimate between category k and category k-1.

The RS model uses the same set of threshold estimates for all sensory attributes or modalities used in the study. A Partial Credit (PC) Model can be used when different labelled category scales are used in the same study (Masters, 1982). The threshold parameter τ_k in Eq. (2) is replaced by τ_{ik}, which indicates a distinct set of threshold estimates for each individual sensory attribute or modality (Bond & Fox, 2015). The log-odds form of the PC model is given by

$$ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \delta_i - \tau_{ik} \tag{3}$$

where:

P_{nik} = probability that assessor (n) rated k for sensory attribute or modality (i),
P_{nik-1} = probability that assessor (n) rated *k-1* for sensory attribute or modality (i),
θ_n = *degree of Leniency* of assessor (n), where n=1,...,a,
δ_i = *degree of Relevance* for sensory attribute (i), where i=1,...,c,
τ_{ik} = boundary estimate between category k and category k-1 for sensory attribute or modality (i)

3.1. The Many-Facet Rasch model

The three unidimensional Rasch models, described in the previous section, have limited applicability in a typical sensory test setting, when the focus is on comparing the differences between more than one product. The Many-Facet Rasch (MFR) Model (Linacre, 1989) is an extension of the dichotomous and polytomous Rasch models that may include one or more additional facets, depending on the objective of the study. An additional parameter (i.e., facet) is included in the MFR model to estimate overall liking for each product. Two assumptions are commonly made when using the 9-point hedonic scale in sensory studies. The first is the assignment of numerical values of 1 to 9 to each category, which assumes that the numerical distance between each category is equally spaced. The second is that assessors would interpret the meaning of each category in roughly the same way regardless of the sensory attribute or modality that is being rated. In this study, we will test the first assumption by selecting a common step structure (McNamara, 1996), whereby the RS model is used for both the Assessor and Criterion facets. The log-odds form for the MFR-RS model, is

$$ln\left[\frac{P_{mnik}}{P_{mnik-1}}\right] = \beta_m - \theta_n - \delta_i - \tau_k \tag{4}$$

where:

P_{mnik} = probability that product(m) being rated k for sensory attribute or modality (i) by assessor (n),
P_{mnik-1} = probability that product(m) being rated *k-1* for sensory attribute or modality (i) by assessor (n),
β_m = *degree of Overall Liking* of product (m), where m=1,...,b,
θ_n = *degree of Criticality* of assessor (n), where n=1,...,a,
δ_i = *degree of Relevance* for sensory characteristic (i), where i=1,...,c,
τ_k = boundary estimate between category k and category k-1.

One common feature of the Rasch Models (Eqs. (2) to (4)) is the inclusion of the item parameter, δ_i. The Individual Attribute Measure (IAM) and Total Attribute Measure (TAM), which have been outlined in Section 2, can be estimated using Eq. (4). The Single Overall Measure (SOM) was estimated using a modified MFR-RS model and the log-odds form can then be expressed as

$$ln\left[\frac{P_{mnk}}{P_{mnk-1}}\right] = \beta_m - \theta_n - \tau_k \tag{5}$$

where:

P_{mnk} = probability that product(m) being rated k by assessor (n),
P_{mnk-1} = probability that product(m) being rated k-*1* by assessor (n),
β_m = *degree of Overall Liking* of product (m), where n=1,...,b,
θ_n = *degree of criticality* of assessor (n), where m=1,...,a,
τ_k = boundary estimate between category k and category k-1.

4. Materials and methods

4.1. Sensory study

The data set was from a study that compared the overall liking of five commercially produced hams (Product 1-5) and five *Bisaro* hams (Product 6-10) that were produced using traditional methods (Ho, Todorov, & Vaz-Velho, 2005). The consumer panel comprised of 56 women and 34 men, aged between 18 and 65 years old, that were staff and students from the College of Technology and Management of the Viana do Castelo Polytechnic Institute. Assessors evaluated 1mm thick slices at room temperature (20-23 °C) under white fluorescent lighting. A balanced incomplete block design, generated using an Optimal Incomplete Cross-Over (OICO) design that was balanced for first order carry-over effects (Périnel & Pagès, 2004), was used to present each assessor with only 4 out of the 10 hams. They were given the first sample to rate, using a paper questionnaire containing a list of 15 descriptors, and were only given the next sample in the presentation order sequence after they had deemed to have finish rating that sample. Samples were evaluated with a list of 15 descriptors in the following order: appearance (Liking of Redness, Liking of Marbling and Liking of Overall Appearance); aroma (Liking of Typical Aroma, Liking of Overall Aroma); texture (Liking of Hardness, Liking of Fibrousness, Liking of Juiciness, Liking of Overall Texture); flavour (Liking of Saltiness, Liking of Sweetness, Liking of Typical Flavour, Liking of Overall Flavour, Liking of Aftertaste) and Overall Liking. They were asked

"Please rate your impressions of the sample, by selecting one of the boxes that matches your impression of liking for each of the following descriptors". The 9-point hedonic scale with 9 labelled categories, from Dislike Extremely, Dislike very much, Dislike slightly, Neither like nor dislike, Like slightly, Like moderately, Like very much to Liked extremely, was used to rate all 15 descriptors.

4.2. Data Analysis

4.2.1. Fitting the Many-Facet Rasch Model

The three measures of Overall Liking, as defined in Section 2, were estimated by fitting a Many-Facet Rasch Rating scale (MFR-RS) model to the sensory data using FACETS (Linacre, 2017a). Eq. (4) was used for estimating the Individual Attribute Measure (IAM) and Total Attribute Measure (TAM), whereas Eq. (5) was used for the Single Overall Measure (SOM). A Joint Maximum Likelihood Estimation (JMLE) method, using default values for FACETS, were used for convergence of the model algorithm (Linacre, 2017a). Two parameters were chosen in FACETS, which allowed the estimates from the three facets to be compared on a common frame of reference on the Many-Facet Wright map. This is firstly done by non-centering the object of measurement and setting the scale origin to a mean of "zero" for other two facets to prevent estimates from being over-constrained (Linacre, 2017a). The Product facet, which was the object of measurement, can then be estimated with respect to the scale origin that was established from centring the Assessor and Criterion Facets. The second parameter that was set was the orientation of the scale for each facet. Only the Product facet was positively orientated (a "+" symbol in brackets) on the Many-Facet Wright Map, whereby a product with a higher Rasch measure (and higher average raw scores) is more liked than another with a lower Rasch measure of the *degree of Overall Liking*. The two other facets were negatively orientated (a "−" symbol in brackets) on the Many-Facet Wright Map. An assessor is more likely to like a product if they have lower Rasch measures (and higher average raw scores) for the *degree of Criticality* than another assessor with a higher Rasch measures. Higher Rasch measures (and lower average raw scores) for individual attributes or sensory modalities on the Criterion facet would indicate that they had a lesser contribution to the measure of Overall Liking. A number of different procedures were subsequently examined after fitting each Many-Facet Rasch model, to ensure that model requirements were met.

4.2.2. Global model fit

The Global fit of the data to the model was firstly examined. An acceptable fit was achieved when there were 5% or less of absolute standardized residuals that were ≥ 2 and about 1% or less that were ≥ 3 (Linacre, 2017a).

4.2.3. Proper functioning of a rating scale

Table 1 shows the guidelines and criteria, as suggested by (Linacre, 2002), that were used to examine the proper functioning of the 9-point hedonic scale. Checking whether the estimates of the rating scale thresholds, also known as Rasch-Andrich thresholds or "step calibrations", are disordered is perhaps one of the most frequently examined guidelines (Tennant & Conaghan, 2007). Linacre (2002) also proposed a minimum distance between thresholds, which depends on the number of categories in the rating scale, for assessing threshold disorder. Rasch-Andrich thresholds, which are the τ_k parameter defined in Eq. (4), can be disordered when there are insufficient observations for a particular rating scale category (Linacre, 2001). When Rasch-Andrich thresholds are disordered, collapsing the number of categories in the rating scale is suggested to improve the overall quality of the measure (Bond & Fox, 2015; Pallant & Tennant, 2007). Models with a revised rating scale, that had a reduced number of categories compared to the original 9-point hedonic scale, were refitted to produce three additional measures of Overall Liking. The global fit and proper functioning of these new measures with the revised rating scales were then re-examined for adequacy.

4.2.4. Criterion, Assessor and Product fit

The OUTFIT mean square (MNSQ), which was used to examine fit of each estimate for individual attributes or sensory modalities (Criterion fit), individual assessors (Assessor fit) and each product (Product fit), is a chi-square statistic that is based on the sum of squared standardised residuals (Bond & Fox, 2015). Four levels for interpreting the fit adequacy have been suggested by Wright and Linacre (1994). OUTFIT MNSQ values between 0.5 and 1.5, which are 'productive for measurement', were considered to indicate adequate fit for all estimates to the MFR-RS models. As the sample size was small, it was important to use as much of the available data for fitting the models. Assessor estimates were also retained when OUTFIT MNSQ values were 'less productive' (OUTFIT < 0.5) or 'unproductive" (OUTFIT between 1.5 to 2.0), but they did 'not to degrade' the measurement. Any estimates with OUTFIT MNSQ values > 2 that 'distort or degrade' the measure were considered for removal, as they misfitted the Rasch model.

4.2.5. Unidimensionality and local independence

The assumptions of unidimensionality and local independence were examined with a Principal Component Analysis of Rasch residuals (PCAR) and by comparing standardized residual correlations respectively (Linacre, 2017d; Smith, 2002). This was done using Winsteps® (Linacre, 2017d) after reformatting the data into two facets, as recommended by Linacre (2017c). Unidimensionality was checked by comparing the disattenuated correlations of the subsets of attributes with the highest (cluster 1) and lowest loadings (cluster 3) on the first unrotated PCAR component, after the Overall Liking dimension had been removed. Disattenuated correlations < 0.57 would likely indicate that the two set of attributes are from different latent variables, whereas values > 0.82 indicates that they probably belong to a single latent variable (Linacre, 2017d). Additionally, the procedure suggested by Smith (2002) was used, whereby unidimensionality holds when no differences are found between two independent estimations of individual person (Assessor) measures from the two subsets of model items (individual attributes or sensory modalities), using a series of t-tests (Hammond et al., 2015; Miller, Slade, Pallant, & Galea, 2010). The trait is unidimensional when the proportion of t-tests from a binomial test is < 5% (or more specifically if the lower bound of the binomial confidence interval is < 5%) (Miller et al., 2010). The assumption of local independence can be tested by comparing person reliability estimates from a model that was fitted after combining pairs of items (attributes or modalities) with a residual correlation >0.3 into single super-items or 'testlets', with the reliability estimates obtained from the model fitted using separate items (Miller et al., 2010; Smith, 2002). A marked reduction in person reliability estimates (Person Separation Index or Cronbach alpha) would indicate a significant local dependency between pairs of items (attributes) (Hagquist, Bruce, & Gustavsson, 2009; Miller et al., 2010).

4.2.6. Statistical analysis

Mean Rasch estimates for each product, assessor and criterion were calculated to determine their relative positions on the unidimensional trait, by representing them onto the Many-Facet version of a Wright Map (Wilson, 2011). For the initial fit for estimating the Rasch measures, the Homogeneity Index was determined to test the hypothesis of equality between Rasch mean estimate (Eckes, 2011). Two Rasch separation statistics were also examined. Reliability (R) is the proportion of the variance of the measures that is not due to measurement error (Eckes, 2011). R in Winsteps® and FACETS is similar to the Person Separation Index (PSI) that is used in Rumm2020 (Andrich, 1982; Wright & Masters, 2002). Strata (S) can

be defined as the number of levels in a measure that are statistically distinctive (Schumacker & Smith, 2007).

Differences between the mean values for each of the 10 products were also examined with parametric and non-parametric analysis of variance (ANOVA) models for the single raw score of overall liking and the six Rasch estimates of Overall Liking. In order to obtain Rasch estimates for all 36 individual replicate evaluations for each of the 10 products, that were required to conduct the above statistical analysis, the MRF-RS models were refitted for all of the six Rasch measures using a different model parametrization in FACETS. The MFR-RS model consisted of an Assessor by Product facet, the Product facet and the Criterion facet. However, in this analysis, the Rasch-Andrich thresholds of the Rating scale and Product facet were anchored using estimates from the initial fit carried out in Section 4.2.1, to ensure that these estimates were equated after refitting the MFR-RS models (Linacre, 2004a, 2017a). Group anchoring was also applied to the *Assessor by Product* facet, by anchoring to the mean estimates of respective products in the Product facet from the analysis in Section 4.2.1., which allowed the mean estimates of each of the 36 estimates for the 10 products to be equated to that of the mean estimates from the first analysis. The Criterion facet was unanchored and was non-centered to prevent the estimates from being over-constrained.

All statistical analyses were conducted using R (R Core Team, 2018). For parametric analyses, the R package *ibd* (Mandal, 2018) was used to conduct an intrablock ANOVA for a balanced incomplete block (BIB) design. Residual analysis was conducted using the *MASS* (Venables & Ripley, 2002) and *car* (Fox & Weisberg, 2011) packages. For non-parametric analyses, *PMCMRplus* (Pohlert, 2018) was used to conduct a Durbin's test for two-way BIB design, followed by the Conover-Iman all-pairs comparison test for a BIB design (Conover & Iman, 1979) with Hochberg's p-adjustment for controlling familywise error (FWE) rates (Hochberg, 1988).

5. Results

5.1. Estimation method and Global Model fit

FACETS uses Joint Maximum Likelihood estimation (JMLE) for estimating the model parameters for the MFR models, which differs from approaches used by other software. For example, marginal maximum likelihood estimation (MMLE) is used in ConQuest (Adams, Wu, & Wilson, 2015), conditional maximum likelihood estimation (CMLE) in eRm (Mair & Hatzinger, 2007) and pairwise conditional estimation (PAIR) in RUMM (Andrich, Lyne, Sheridan, & Luo, 2003).

Eckes (2011) noted that there has been much debate as to which of these methods should be preferred, as some proponents have suggested that JMLE is mildly inconsistent. Other estimation methods, such as CMLE, can also be inconsistent under certain conditions (Linacre, 2004b). Consistency, has been defined by Linacre (2004b), as 'the property that, given an infinite amount of data which fit the model, the estimation procedure would recover the values of the parameters used to generate those values'. Eckes (2011) also noted some authors have suggested that there might be no practical significant difference between the use of different estimation methods. Since the aim of the study was not to compare differences in consistency or estimation bias of different estimation approaches, henceforth, we shall only consider issues related to fit that is based solely on estimates produced by FACETS.

Various methods have been proposed for assessing model fit (Fischer & Molenaar, 1995). However, many of these methods can only be used with the specific estimation methods previously outlined. Therefore, different fit statistics are used in different Rasch software, depending on the estimation methods implemented. The method of comparing the absolute standardized residuals (Eckes, 2011), that has been described in Section 4.2.2., has been used by some authors for assessing the global model fit for MFR models using FACETS (Eckes, 2005; Toffoli, Andrade, & Bornia, 2016). The three original Rasch models and the models with their revised scales all showed acceptable model fit, when comparing their absolute standardized residuals to these criteria (Table 2).

5.2. Proper functioning of the 9-point hedonic scale

Table 3 shows that the 9-point hedonic scale, for the Individual Attribute Measure (IAM-R9) and the Single Overall Measure (SOM), both functioned within the specified criteria for all the metrics that were considered to be essential for stability of the measure, fit accuracy, sample description and inference (Table 1). However, the Total Attribute Measure (TAM) did not meet all those essential criteria. The observed average measure did not increase monotonically up the scale between "Dislike Extremely" and "Dislike very much".

The Rasch-Andrich thresholds did not increase monotonically up the scale for middle categories of the 9-point hedonic scale for all three Rasch measures. This was between "Dislike moderately" and "Dislike slightly" and between "Neither like nor dislike" and "Liked slightly", for IAM-R9 and TAM-R9, and only between "Neither like nor dislike" and "Liked slightly" for SOM-R9. Additionally, the minimum distance between the Rasch-Andrich thresholds for some of the categories were smaller than the recommended value of 0.45 for a rating scale

with 9 categories for all three Rasch measures (Table 1). Acceptable values for these two Rasch-Andrich thresholds metrics were achieved by combining some of the adjacent categories of the 9-point hedonic scale, as suggested by (Bond & Fox, 2015). The IAM-R6 was estimated with a revised 6-category scale, with the category "Dislike very much" combined with "Dislike moderately", "Dislike slightly" with "Neither like nor dislike" and "Liked slightly" with "Liked moderately". Similarly, as shown in (Table 4), TAM-R5 was estimated using a revised 5-category and SOM-R4 with a revised 4-category scale.

5.3. How well do Criteria, Panellists and Products fit the Rasch model?

The fit adequacy of each criterion (ie., sensory attributes or modalities) and assessor to the MFR models must be examined before we can use the product Rasch measures as a means of comparing any differences between the Overall Liking of the 10 products. For the Criterion facet, a two-stage procedure is adopted to examine fit, where individual estimates are firstly examined and any misfitting criteria should be removed. The model should then be refitted and the whole process of checking model fit begins again. Unidimensionality and local independence assumptions are then examined, as described in the next section. None of the sensory attributes or modalities were found to misfit the IAM and TAM models with the original or revised scales (Table 2).

An assessor that misfits the Rasch model might influence the criterion and product estimates. Therefore, identifying misfitting assessors are important and a decision has to be made as to either remove all of the estimates belonging to that person or to only remove observations that have been identified as misfitting from that person (Boone, Staver, & Yale, 2014). A small percentage of assessors were found to have misfitted all of the Rasch models, as shown in Table 2. However, a few misfitting assessors are unlikely to have a significant effect on criterion and product estimates (Wright & Linacre, 1994). Since the inclusion of these misfitting assessors did not seem to have any effect on the fit for Criteria or Products, none of them were removed.

5.4. Are the Overall Liking Rasch measures unidimensional measures?

The main assumption made in developing the Overall Liking Rasch measure was that a unidimensional measure could be constructed from a set of individual attributes or sensory modalities. A consumer's OVERALL LIKING of a food product would then be estimated by taking into consideration the extent to which each individually chosen attribute or modality contributed to that Overall Liking

measure. The assumption of unidimensionality and local independence of criteria were considered for the four Rasch models that had more than one criterion.

IAM-R9 was found to be unidimensional, as the lower bound of the binomial confidence interval was below 5% for (Table 2). However, IAM-R6 had a lower bound value of 5.01%. However, the disattenuated correlation was 0.63 for both measures between the highest and lowest loaded clusters on the first PCAR component. This value was higher than the value of 0.57, which was suggested by (Linacre, 2017d), if the measure would have likely been multidimensional. Linacre (2017d) proposed that an eigenvalue of less than 3 for the unexplained variance on the first unrotated PCAR component would probably suggest unidimensionality. The eigenvalues were 2.42 for the IAM-R6 and 2.58 for the IAM-R9. It might be reasonable to assume that the IAM-R6 was most likely to be unidimensional like that of the IAM-R9, based solely on the similarities in the values for the disattenuated correlation and eigenvalue for the unexplained variance on the first unrotated PCAR component. Hence, both IAM measures were most likely unidimensional.

TAM-R9 and TAM-R5 were both found to be unidimensional, as the lower bound of the binomial confidence interval was below 5% for (Table 2). However, it is not clear if the small number of items might have affected the test of unidimensionality suggested by Smith (2002). Overall Flavour and Overall Texture had the two highest positive loadings on the first PCAR component, with Overall Appearance and Overall Aroma showing the lowest negative loadings. The disattenuated correlation between these two clusters was 0.817, which was very close to the value (i.e., 0.82), that would indicate that the items (modalities) belonged to a single latent variable (Linacre, 2017d). The eigenvalues for the unexplained variance on the first unrotated PCAR component for both measures were less than 3, at 1.85 for the TAM-R5 and 1.95 for the TAM-R9, indicating that the measures were probably unidimensional. No problems were found for local independence, as there were no standardized residuals correlations for either model ≥ 0.3.

5.5. Rasch Overall Liking measures as a representation of the construct of OVERALL LIKING

Fig. 3 shows a Many-Facet Wright Map that represents estimates for each of the three facets for the IAM-R9 and a fourth 'Scale' column indicating how the 9-point hedonic scale is mapped onto the logit scale. The horizontal lines indicate Rasch-half-point thresholds that represent the 'average score half-way between two adjacent categories', as opposed to the Rasch-Andrich thresholds and the Rasch-Thurstone thresholds that examine 'which of two adjacent categories is

more likely to be observed?' and 'which category is more likely to be observed?' respectively (Linacre, 2006).

The MFR-RS model indicated that the distance between each category label in the 9-point hedonic scale is unequally spaced, with that gap between two adjacent categories getting larger the further away from the centre of the scale. The results show how the MFR-RS model might be used to indicate actual perceived OVERALL LIKING, not normally associated with category scales (Synder & Bartoshuk, 2015). Collapsing the scale to produce a revised scale still preserves, although to a lesser extent, the property of unequally spaced categories that get larger the further away from the centre of the scale (Fig. 4). This property can also be found in category-ratio scales like the LAM, LHS or the hedonic gLMS (Bartoshuk et al., 2005; Lim et al., 2009; Schutz & Cardello, 2001). However, unlike category-ratio scales where a category label indicates a single point along the scale, each category label in the Rasch measure covers a greater portion of the scale, as shown in Fig. 3. One advantage using the MFR model with 9-point hedonic scale, is that it only requires an assessor to choose a distinct category label that represents their rating of liking. This is likely to reduce the variability in responses compared to using category-ratio scales, that require the assessor to select an arbitrary point along a continuous line. The raw scores, which are obtained from assigning numbers from 1 to 9 to each labelled category, are ordinal. These raw scores are then converted to interval-scale data by using an appropriate MFR model, which has been shown to be represented on an interval scale that is similar to category-ratio scales.

Another advantage of using the Many-Facet Wright Map, is that it allows us to visualise the importance of each sensory attribute or sensory modality in the estimation of Overall Liking for IAM and TAM. The Criterion facet in Fig. 3 indicates that texture attributes, like Hardness and Juiciness contribute more to the measure of Overall Liking than Sweetness, that contributed the least. The hams were significantly saltier than sweet. The position of the estimates of the sensory modalities in TAM were much more centred and less dispersed compared to the sensory attributes in IAM, with Overall Texture having a slightly stronger contribution that Overall Flavour.

5.6. Were there any differences in OVERALL LIKING between products?

The homogeneity test found that there were significant differences ($p<0.01$) between the Rasch mean estimates for the 10 products with the two Rasch measures with the sensory attributes and with sensory modalities from the initial fit of the MFR-RS models (Table 5). Reliability and Strata values for IAM-R9

and IAM-R6 indicated that the 10 products could be separated into at least three groups with TAM-R9 and into just two groups and TAM-R5 at a 5% significance level (Fisher, 1992, 2008). The measures with a single rating of overall liking (i.e, SOM-R9 and SOM-R4) were unlikely to find any significant differences between the 10 products. The small standard errors that were estimated, shown in brackets in Table 6, could explain why significant differences were found for the initial fit of these models.

The MFR-RS models were refitted to obtain 36 individual estimates for each of the 10 products, as described in Section 4.2.6, so that further comparisons could be made between the raw scores for the single estimate of Overall Liking and the six Rasch measures. Standard errors for each sample were much larger for all Rasch measures, except for SOM-R4 (Table 6), due to differences in the number of observations used in the estimation process between the initial and refitted models. For example, the average standard errors of 0.04 logits for the means of each of the 10 products for IAM-R9 were based on an average of 354 observations. However, an average of only just 10 observations were used in estimating each of the 36 estimates for every product.

An intrablock BIB ANOVA showed no significant differences between the 10 products for all of the six Rasch measures and the raw scores (Table 5). However, a Durbin test was also conducted as residual analysis indicated that non-normality of residuals and outliers were found for IAM-R9 and TAM-R9, while residuals from SOM-R9 and SOM-R6 were found to be heteroscedastic. Rasch estimates in small sample sizes have been found to be non-normal (Biehler, Holling, & Doebler, 2015; Doebler, Doebler, & Holling, 2013; Klauer, 1991). However, collapsing the original scale in these two Rasch measures produced estimates with standardized residuals that were found to be normally distributed and homoscedastic.

The Durbin test confirmed the results obtained by the intrablock BIB analysis for all of the measures, except for SOM-R9 ($p<0.01$). However, no significant differences were found at a 5% level between the mean estimates of any of the 10 products after applying Hochberg's adjustment method for pairwise comparison (Hochberg, 1988).

6. Discussion

The Many-Facet Rasch Rating scale (MFR-RS) model has been used to produce interval-scaled estimates of Overall Liking. The extremely small differences between product means compared to their standard errors, resulted in no significant differences between the mean Overall Liking of products. However, more

precise mean estimates (i.e., smaller standard errors) were produced with the IAM and TAM Rasch measures, compared to means for the Rasch measures (i.e., SOM) or the raw score estimate with a single holistic variable of overall liking. The use of 10 sensory attributes and the 4 sensory modalities effectively increased the number of observations used to estimate the Rasch parameters in the model, producing tighter standard errors. Collapsing the original 9-point scale is recommended to reduce the problems associated with non-normality and heteroscedasticity, if parametric statistical methods are to be used.

One of the main advantages of using Many-Facet Rasch models is the potential assessment of invariance in sensory measurement. Engelhard and Wind (2018) proposed five requirements for invariant measurement for raters that could be adapted for use in sensory measurement with assessors:

1. **Assessor invariant measurement of products**. Product evaluations must be independent from the sensory panel. In the context of the measurement of OVERALL LIKING, we should expect any random sample of assessors with similar characteristics (or preferences) from a population to evaluate the products in a similar manner, which would result in similar estimates of Overall Liking. *Differential facet functioning (DFF* can be used to test if significant differences exists between how assessors rate products with different characteristics across all sensory criteria (Engelhard, 2013).

2. **Assessor invariant calibration of sensory criteria onto a sensory construct**. The calibration of sensory criteria used to define the sensory construct should be independent of the particular assessors used in the product evaluation. *Differential item functioning (DIF)* can be used to determine whether sensory criteria are invariant across different types of assessors. If a sensory criterion is found to have *DIF*, then this would indicate that it is bias (Engelhard, 2013). For example, *DIF* could exist if a trained panel consisted of assessors with genetic differences that affected their ability to rate certain sensory characteristics in a different manner.

3. **Assessor invariant calibration of rating scales**. The structure of the sensory rating scale should be independent of the assessors used in the product evaluation. The 9-point hedonic scale produced unequally spaced scale categories, whereby the distance between each category increased the further away from the central category. In a recent presentation at Eurosense 2018, I applied two different Many-Facet Rach models that used a combination of the Rating Scale model and Partial Credit model to examine (i) if the

rating scale structure of the 9-point hedonic scale was comparable between sensory attributes from the IAM and (ii) if the rating scale was perceived differently between assessors (Ho, 2018). Differences in scale usage across attributes and assessors had a negligible effect on the mean estimates of Overall Liking for the 10 products. Full details of the results of this study will be a subject of a future publication.

4. **Invariant location of Assessors**. The *degree of Criticality* of an assessor should be invariant to the product, sensory criteria and rating scale that is used. For example, we would expect a trained assessor to always rate a product that has a higher intensity at a higher level of intensity than a product with a lower inherent intensity. Differential rater functioning (DRF) has been used to examine rater drift over time (Myford & Wolfe, 2009). Context effects, biases and classical psychological errors in sensory measurement (Lawless & Heymann, 2010) can detected and measured with Many-Facet Rasch models (Myford & Wolfe, 2003, 2004).

5. **Assessor invariant Wright map**. All assessors should have a common understanding and use of the assessment procedures, that subsequently allows for assessors, sensory criteria and products to be simultaneously located on an underlying continuous latent sensory construct.

This study introduces the concept of construct modelling and the use of the Many-Facet Rasch Models for measuring sensory constructs like OVERALL LIKING. Advantages of its use and parametric and non-parametric analysis of variance models were used to compare differences between mean estimates of raw scores from a single holistic variable with ordinal raw scores and Rasch measures for 10 ham products. Future studies will explore the use of Many-Facet Rasch models in measuring invariance in sensory measurement.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised item response modelling software [Computer software] Version 4.* Camberwell, Victoria: Australian Council for Educational Research.

Andersen, B., Brockhoff, P. B., & Hyldig, G. (2019). The importance of liking of appearance, -odour, -taste and -texture in the evaluation of overall liking. a comparison with the evaluation of sensory satisfaction. *Food Quality and Preference*, *71*, 228–232.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–574.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95–104.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). *Rasch unitary measurement model 2020 (rumm2020)*. Perth, Australia. 15.: RUMM Laboratory Pty.

Bakke, A., & Vickers, Z. (2007). Consumer liking of refined and whole wheat breads. *Journal of Food Science*, *72*(7), S473–S480.

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.

Bayarri, S., Carbonell, I., Barrios, E. X., & Costell, E. (2010). Acceptability of yogurt and yogurt-like products: Influence of product information and consumer characteristics and preferences. *Journal of Sensory Studies*, *25*, 171–189.

Biehler, M., Holling, H., & Doebler, P. (2015). Saddlepoint approximations of the distribution of the person parameter in the two parameter logistic model. *Psychometrika*, *80*(3), 665-688.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model* (Third ed.). New York: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Netherlands: Springer.

Conover, W. J., & Iman, R. L. (1979). *On multiple-comparisons procedures* (Tech. Rep.). LA-7677-MS. Los Alamos Scientific Laboratory.

Doebler, A., Doebler, P., & Holling, H. (2013). Optimal and most exact confidence intervals for person parameters in item response theory models. *Psychometrika*, *78*(1), 98-115.

Drake, M. A., & Gerard, P. D. (2003). Consumer attitudes and acceptability of soy-fortified yogurts. *Journal of Food Science*, *68*(3), 1118-1122.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A Many-Facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221.

Eckes, T. (2011). *Introduction to Many-Facet Rasch measurement* (Vol. 22). Frankfurt: Peter Lang.

El Dine, N. A., & Olabi, A. (2009). Effect of reference foods in repeated acceptability tests: Testing familiar and novel foods using 2 acceptability scales.

Journal of Food Science, *74*(2), S97–S106.

Engelhard, G. (2013). *Invariant measurement: using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales*. New York: Routledge.

Fischer, G., & Molenaar, I. (1995). *Rasch Models: Foundations, recent developments and applications*. New York: Springer-Verlag.

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions*, *6*(3), 238.

Fisher, W. P. (2008). The cash value of reliability. *Rasch Measurement Transactions*, *22*(1), 1160-1163.

Forde, C., & Delahunty, C. (2004). Understanding the role cross-modal sensory interactions play in food acceptability in younger and older consumers. *Food Quality and Preference*, *15*(7), 715–727.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second ed.). Thousand Oaks CA: Sage.

Hagquist, C., Bruce, M., & Gustavsson, P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*, *46*(3), 380–393.

Hammond, A., Tennant, A., Tyson, S. F., Nordenskiöld, U., Hawkins, R., & Prior, Y. (2015). The reliability and validity of the english version of the evaluation of daily activity questionnaire for people with rheumatoid arthritis. *Rheumatology*, *54*(9), 1605-1615.

Ho, P. (2018). Rethinking hedonic scaling: A new approach to analysing the 9-point hedonic scale with Rasch modelling. In *Eurosense 2018. Eight European Conference on Sensory and Consumer Research.* 2-5 September 2018, Verona, Italy..

Ho, P., Todorov, S. D., & Vaz-Velho, M. (2005). *A comparison of the overall acceptability between industrially and traditionally produced portuguese smoked ham.* SAAFost Biennial Congress, Stellenbosch, South Africa, 5-7 September 2005.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(3), 800–803.

Hough, G., & Sánchez, R. (1998). Descriptive analysis and external preference mapping of powdered chocolate milk. *Food Quality and Preference*, *9*(4), 197–204.

Klauer, K. C. (1991). Exact and best confidence intervals for the ability parameter of the rasch model. *Psychometrika*, *56*(3), 535-547.

Lawless, H. T., Cardello, A. V., Chapman, K. W., Lesher, L. L., Given, Z., & Schutz, H. G. (2010). A comparison of the effectiveness of hedonic scales and end-anchor compression effects. *Journal of Sensory Studies*, *25*(s1), 18–34.

Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food* (2nd ed.). New York: Springer.

Lawless, H. T., Popper, R., & Kroll, B. J. (2010). A comparison of the labeled magnitude (LAM) scale, an 11-point category scale and the traditional 9-point hedonic scale. *Food Quality and Preference*, *21*(1), 4–12.

Lawless, H. T., Sinopoli, D., & Chapman, K. W. (2010). A comparison of the labeled affective magnitude scale and the 9-point hedonic scale and examination of categorical behavior. *Journal of Sensory Studies*, *25*(s1), 54–66.

Lim, J., Wood, A., & Green, B. G. (2009). Derivation and evaluation of a labeled hedonic scale. *Chemical Senses*, *34*(9), 739–751.

Linacre, J. M. (1989). *Many-Facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2001). Category, step and threshold: Definitions & disordering. *Rasch Measurement Transactions*, *15*(1), 794.

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

Linacre, J. M. (2004a). Ability estimated from adding item difficulties. *Rasch Measurement Transactions*, *18*(3), 993.

Linacre, J. M. (2004b). Rasch model estimation: Further topics. *Journal of Applied Measurement*, *5*(1), 95–110.

Linacre, J. M. (2006). Demarcating category intervals. *Rasch Measurement Transactions*, *19*(3), 10341–10343.

Linacre, J. M. (2008). The expected value of a point-biserial (or similar) correlation. *Rasch Measurement Transactions*, *22*(1), 1154.

Linacre, J. M. (2017a). *Facets computer program for many-facet rasch measurement*. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2017b). *Step intervals for rating scale categories when using PCM [Online forum comment].* Retrieved 14 July 2017, from http://raschforum.boards.net/thread/660/intervals-rating-scale-categories-using

Linacre, J. M. (2017c). *Testing unidimensionality in facets. [Online forum comment].* Retrieved 12 December 2017, from http://raschforum.boards.net/thread/288/testing-unidimensionality-facets

Linacre, J. M. (2017d). *Winsteps® Rasch measurement computer program user's guide.* Beaverton, Oregon: Winsteps.com.

Mair, P., & Hatzinger, R. (2007). Extended rasch modeling: The erm package for the application of irt models in r. *Journal of Statistical Software*, *20*(9), 1–20.

Mandal, B. N. (2018). ibd: Incomplete block designs [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ibd (R package version 1.4)

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Miller, K. J., Slade, A. L., Pallant, J. F., & Galea, M. P. (2010). Evaluation of the psychometric properties of the upper limb subscales of the motor assessment scale using a Rasch analysis model. *Journal of Rehabilitation Medicine*, *42*(4), 315–322.

Moskowitz, H., & Krieger, B. (1993). What sensory characteristics drive product quality? An assessment of individual differences. *Journal of Sensory Studies*, *8*, 271–282.

Moskowitz, H., & Krieger, B. (1995). The contribution of sensory liking to overall liking: an analysis of six food categories. *Food Quality and Preference*, *6*, 83–90.

Mudgil, D., Baraka, S., & Khatkar, B. S. (2017). Cookie texture, spread ratio and sensory acceptability of cookies as a function of soluble dietary fiber, baking time and different water levels. *LWT-Food Science and Technology*, *80*, 537–542.

Myford, C., & Wolfe, E. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement : Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C., & Wolfe, E. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement : Part II. *Journal of Applied Measurement*, *5*(2), 386-422.

Myford, C., & Wolfe, E. (2009). Monitoring rater performance over time: A framework ofr detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389.

Nicolas, L., Marquilly, C., & O'Mahony, M. (2010). The 9-point hedonic scale: Are words and numbers compatible? *Food Quality and Preference*, *21*(8), 1008–1015.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the hospital anxiety and depression scale

(HADS). *British Journal of Clinical Psychology*, *46*, 1–18.

Périnel, E., & Pagès, J. (2004). Optimal nested cross-over designs in sensory analysis. *Food Quality and Preference*, *15*(5), 439–446.

Peryam, D. R., & Girardot, N. F. (1952). Advanced taste-test method. *Food Engineering*, *24*(7), 58–61.

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

Pham, A. J., Schilling, M. W., Mikel, W. B., Williams, J. B., Martin, J. M., & Coggins, P. C. (2008). Relationships between sensory descriptors, consumer acceptability and volatile flavor compounds of american dry-cured ham. *Meat Science*, *80*(3), 728–737.

Pohlert, T. (2018). Pmcmrplus: Calculate pairwise multiple comparisons of mean rank sums extended [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=PMCMRplus (R package version 1.4.0)

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ramcharitar, A., Badrie, N., Marrfeldt-Beman, M., Matsuo, H., & Ridley, C. (2005). Consumer acceptability of muffins with flaxseed (Linum usitatissimum). *Journal of Food Science*, *70*(7), S504–S507.

Rankin, L. D., Fada, R. D., & Bingham, M. (2000). Acceptability of oatmeal chocolate chip cookies prepared using pureed white beans as a fat ingredient substitute. *Journal of the American Dietetic Association*, *100*(7), 831—833.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Schumacker, R. E., & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement*, *67*(3), 394–409.

Schutz, H. G., & Cardello, A. V. (2001). A labeled affective magnitude LAM scale for assessing food liking/disliking. *Journal of Sensory Studies*, *16*(2), 117–159.

Shao, C. H., Avens, J. S., Schmidt, G. R., & Maga, J. A. (1999). Functional, sensory, and microbiological properties of restructured beef and emu steaks. *Journal of Food Science*, *64*(6), 1052–1054.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals.

Journal of Applied Measurement, *3*(2), 147–163.

Synder, D. J., & Bartoshuk, L. M. (2015). Psychophysical measurement of human oral experience. In A. R. Hirsch (Ed.), *Nutrition and sensation* (pp. 103–138). CRC Press.

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 & Rheumatism*, *57*(8), 1358–1362.

Toffoli, S., Andrade, D., & Bornia, A. (2016). Evaluation of open items using the many-facet rasch model. *Journal of Applied Statistics*, *43*(2), 299–316.

Varela, P., Beltrán, J., & Fiszman, S. (2014). An alternative way to uncover drivers of coffee liking: Preference mapping based on consumers' preference ranking and open comments. *Food Quality and Preference*, *32*, 152–159.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer.

Villanueva, N. D., Petenate, A. J., & Da Silva, M. A. (2000). Performance of three affective methods and diagnosis of the anova model. *Food Quality and Preference*, *11*(5), 363–370.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. New York: Psychology Press.

Wilson, M. (2011). Some notes on the term: "Wright Map". *Rasch Measurement Transactions*, *25*(3), 1331.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions*, *16*(3), 888.

Young, N. D., Drake, M., Lopetcharat, K., & McDaniel, M. R. (2004). Preference mapping of cheddar cheese with varying maturity levels. *Journal of Dairy Science*, *87*(1), 11—19.

Table 1
Guidelines used for assessing proper functioning of a rating scale[1]

Metric	Description
Item-level indices of polarity	Items should be oriented in the same direction as the latent variable, as it is essential for measure stability, fit accuracy, sample description and inference. Point-biserial or point-measures correlations[2] can be used to identify items that have reversed polarities, when both positively and negatively-orientated items are used.
Category frequency	At least 10 observations of each category are helpful for fit accuracy and sample inference but essential for measure stability.
Observation distribution	Uniform distribution of observations are useful for optimal step calibration and when considering collapsing categories. Helpful for measure stability and sample inference.
Average measures	Values should increase monotonically up the scale categories, as it helpful for measure stability but essential for fit accuracy, sample description and inference.
OUTFIT Mean Squares	Values less than 2.0 are helpful for measure stability, sample description and inference but essential for fit accuracy.
Threshold calibrations	Values increase monotonically up the scale categories (i.e. ordered thresholds). Helpful for sample inference.
Minimum distance between threshold	0.45, 0.7, 0.81 for a 9-,6- and 5- category scale respectively[3]. Helpful for sample inference.
Maximum distance between threshold	No larger than 5 logits to avoid gaps in the variable. Helpful for measure stability.

[1] Bond and Fox (2015); Linacre (2002)
[2] Linacre (2008, 2017d)
[3] Central distance using $ln(x/(m - x + 1)$ for x=1,....m, where m=n-1 for a n-category scale.(Linacre, 2017b)

Table 2
Summary fit statistics from the initial fit for different Rasch models

	Global fit[1]			$OUTFIT_C$[3]	$OUTFIT_A$[3]		$OUTFIT_P$[3]	UniDimenionality
Model	% Stdres	% Stdres	Total[2]	% Fit	% Fit	% Misfit	% Fit	% t-test(CL%)
Original Scale								
IAM-R9	3.7(130)	0.7(24)	3540	100	82	6	100	1.1(0.3)
TAM-R9	3.2(45)	0.2(6)	1405	100	78	4	100	1.7(0.6)
SOM-R9	3.6(13)	0.3(1)	360	100	63	4	100	
Revised Scale								
IAM-R6	4.7(166)	0.2(6)	3540	100	78	4	100	7.5(5.0)
TAM-R5	4.1(57)	0.1(2)	1405	100	77	6	100	0.2(<0.01)
SOM-R4	2.8(10)	0	360	100	58	7	100	
Criteria	≤ 5%	≤ 1%		0.5-1.5	0.5-1.5	>2.0	0.5-1.5	CL< 5%

[1] Percentage and number of observations (in brackets) of absolute standardized residuals
[2] Total Number of observations
[3] Unweighted mean squares for Criteria($OUTFIT_C$), Assessor ($OUTFIT_A$) and Product($OUTFIT_P$)

Table 3
Category statistics for the 9-point hedonic scale for different MFR-RS models

			Average Measure			Rasch-Andrich Threshold	
Category	Label	Counts[1]	Observed[2]	Expected[3]	OUTFIT[4]	Measure[5]	| Difference |[6]
IAM-R9							
1	Dislike extremely	21 (0.6)	-0.30	-0.29	0.9		
2	Dislike very much	141 (4.0)	-0.20	-0.20	1.1	-2.15	
3	Dislike moderately	184 (5.2)	-0.11	-0.11	1.0	-0.42	1.73
4	Dislike slightly	453 (12.8)	0.02	-0.01	1.1	-0.96*	0.54
5	Neither like nor dislike	486 (13.7)	0.10	0.10	1.0	-0.03	0.93
6	Like slightly	761 (21.5)	0.19	0.23	1.0	-0.28*	0.25**
7	Like moderately	763 (21.6)	0.39	0.38	0.9	0.30	0.58
8	Like very much	663 (18.7)	0.57	0.56	1.0	0.61	0.31**
9	Like extremely	68 (1.9)	0.82	0.76	1.0	2.94	2.33
TAM-R9							
1	Dislike extremely	4 (0.3)	-0.18	-0.30	1.2		
2	Dislike very much	47 (3.3)	-0.30*	-0.23	0.9	-2.73	
3	Dislike moderately	69 (4.9)	-0.16	-0.14	0.9	-0.56	2.17
4	Dislike slightly	157 (11.2)	0.00	-0.03	1.1	-0.91*	0.35**
5	Neither like nor dislike	186 (13.2)	0.14	0.10	1.1	-0.14	0.77
6	Like slightly	324 (23.1)	0.22	0.26	1.0	-0.38*	0.24**
7	Like moderately	336 (23.9)	0.48	0.47	1.0	0.32	0.7
8	Like very much	262 (18.6)	0.76	0.75	1.0	0.85	0.53
9	Like extremely	20 (1.4)	1.22	1.22	1.0	3.54	2.69
SOM-R9							
1	Dislike extremely	0 (0)					
2	Dislike very much	10 (2.8)	-1.08	-0.79	0.6		
3	Dislike moderately	22 (6.1)	-0.64	-0.63	0.9	-1.5	
4	Dislike slightly	40 (11.1)	-0.41	-0.48	1.1	-1.15	0.35**
5	Neither like nor dislike	51 (14.2)	-0.31	-0.33	0.8	-0.65	0.5
6	Like slightly	80 (22.2)	-0.14	-0.16	1.4	-0.7*	0.05**
7	Like moderately	75 (20.8)	0.12	0.09	0.9	0.02	0.72
8	Like very much	78 (21.7)	0.41	0.5	0.5	0.23	0.21**
9	Like extremely	4 (1.1)	1.67	1.09	0.9	3.75	3.52

[1] Number and percentage (shown in brackets) of observations used in each category
[2] Modelled average measure in log-odds units (logits)
[3] Expected average measure if data fitted the Rasch model
[4] Unweighted mean square for observations in each category
[5] Location on the latent variable, relative to the centre of the scale, where adjacent categories are equally probable
[6] Absolute difference between Rasch-Andrich threshold values of two adjacent categories
* Measure value that does not increase with a higher rating category
** Difference is smaller than minimum acceptable threshold value (0.45) for a 9-point hedonic scale

Table 4
Category statistics for different MFR-RS models after revision of scale categories

			Average Measure			Rasch-Andrich Threshold	
Category	Label	Counts[1]	Observed[2]	Expected[3]	OUTFIT[4]	Measure[5]	| Difference |[6]
IAM-R6							
1	Dislike extremely	21 (0.6)	-0.68	-0.69	1.0		
2	Dislike very much/ Dislike moderately	325 (9.2)	-0.37	-0.37	1.0	-3.27	
3	Dislike slightly/ Neither like nor dislike	939 (26.5)	0.00	-0.01	1.0	-1.25	2.02
4	Like slightly/ Like moderately	1524 (43.1)	0.38	0.39	1.0	-0.3	0.95
5	Like very much	663 (18.7)	0.86	0.85	1.0	1.45	1.75
6	Like extremely	68 (1.9)	1.40	1.32	1.0	3.37	1.92
TAM-R5							
1	Dislike extremely/ Dislike very much/ Dislike moderately	120 (8.5)	-1.5	-1.43	0.9		
2	Dislike slightly/ Neither like nor dislike	343 (24.4)	-0.98	-1.04	1.1	-2.3	
3	Like slightly/ Like moderately	660 (47.0)	-0.55	-0.53	1.0	-1.45	0.85
4	Like very much	262 (18.6)	0.12	0.10	1.0	0.70	2.15
5	Like extremely	20 (1.4)	0.89	0.88	1.0	3.05	2.35
SOM-R4							
1	Dislike extremely/ Dislike very much/ Dislike moderately	32 (9.0)	-0.77	-0.57	0.8		
2	Dislike slightly/ Neither like nor dislike	91 (25.6)	0.05	-0.03	1.0	-1.35	
3	Like slightly/ Like moderately	155 (43.5)	0.58	0.56	1.1	-0.27	1.08
4	Like very much/ Like extremely	78 (21.9)	1.30	1.35	1.1	1.62	1.89

[1] Number and percentage (shown in brackets) of observations used in each category
[2] Modelled average measure in log-odds units (logits)
[3] Expected average measure if data fitted the Rasch model
[4] Unweighted mean square for observations in each category
[5] Location on the latent variable, relative to the centre of the scale, where adjacent categories are equally probable
[6] Absolute difference between Rasch-Andrich threshold values of two adjacent categories

Table 5
Comparison between Homogeneity index and Rasch separation statistics with an intrablock BIB analysis and the Durbin test for Raw scores and different Rasch measures (in logits)

		Original Scale			Collapsed Scale		
Test	Raw	IAM-R9	TAM-R9	SOM-R9	IAM-R6	TAM-R5	SOM-R4
Homogeneity index[1]							
χ^2		75	35.4	11.9	63.0	31.0	10.5
p-value		< 0.01	< 0.01	0.22	< 0.01	< 0.01	0.31
Separation[1]							
Reliability		0.89	0.77	0.26	0.86	0.72	0.15
Strata		4.04	2.80	1.13	3.62	2.47	0.89
Intrablock BIB[2]							
$F_{Product}$	0.83	1.12	0.71	0.58	1.32	1.08	0.95
p-value	0.5907	0.3465	0.6993	0.8094	0.2260	0.3745	0.4821
Residual Analysis[2]							
Shapiro-Wilk	0.1125	0.011	0.009	0.04321	0.3562	0.4481	0.1116
Brown-Forsythe	0.1646	0.1925	0.1689	0.03235	0.4388	0.2012	0.03234
Bonferroni Outlier	>1	0.0279	0.007	0.9258	0.1012	0.05224	0.5970
Durbin test[2]							
χ^2	6.82	10.24	3.03	9.44	1.28	6.58	22.88
p-value	0.6559	0.3314	0.9630	0.3977	0.2573	0.6803	0.006

[1] Based on estimates from initial fit for the different MFR-RS models
[2] Based on estimates from refit for the different MFR-RS models

Table 6
Mean and standard errors[1] for Raw scores and different Rasch measures (in logits) for each Product

		Original Scale			Collapsed Scale		
Product	Raw	IAM-R9	TAM-R9	SOM-R9	IAM-R6	TAM-R5	SOM-R4
1	6.00±0.31	0.32±0.14 (0.04)	0.40±0.19 (0.07)	-0.09±0.26 (0.14)	0.46±0.21 (0.07)	-0.42±0.27 (0.13)	0.67±0.26 (0.25)
2	6.47±0.26	0.44±0.14 (0.04)	0.61±0.17 (0.07)	0.19±0.24 (0.17)	0.62±0.18 (0.07)	-0.15±0.21 (0.12)	0.77±0.23 (0.30)
3	5.72±0.28	0.10±0.11 (0.03)	0.20±0.15 (0.07)	-0.25±0.20 (0.12)	0.04±0.17 (0.06)	-0.79±0.23 (0.11)	0.21±0.25 (0.23)
4	6.11±0.25	0.19±0.09 (0.04)	0.28±0.13 (0.06)	0.00±0.20 (0.13)	0.20±0.14 (0.07)	-0.71±0.21 (0.11)	0.61±0.23 (0.23)
5	6.53±0.22	0.37±0.11 (0.04)	0.45±0.14 (0.06)	0.09±0.19 (0.14)	0.49±0.16 (0.07)	-0.40±0.19 (0.12)	0.81±0.22 (0.24)
6	5.94±0.28	0.23±0.13 (0.04)	0.26±0.17 (0.07)	-0.16±0.23 (0.12)	0.27±0.20 (0.07)	-0.68±0.25 (0.11)	0.28±0.25 (0.24)
7	5.86±0.30	0.22±0.12 (0.04)	0.22±0.15 (0.06)	-0.11±0.24 (0.14)	0.27±0.20 (0.07)	-0.69±0.26 (0.12)	0.26±0.27 (0.26)
8	6.06±0.23	0.28±0.11 (0.04)	0.32±0.15 (0.06)	0.08±0.18 (0.13)	0.38±0.17 (0.07)	-0.61±0.22 (0.11)	0.63±0.20 (0.23)
9	5.53±0.33	0.13±0.15 (0.03)	0.27±0.19 (0.06)	-0.25±0.24 (0.12)	0.20±0.23 (0.07)	-0.60±0.30 (0.11)	0.14±0.30 (0.24)
10	5.94±0.28	0.17±0.12 (0.04)	0.21±0.12 (0.06)	-0.09±0.21 (0.12)	0.17±0.19 (0.07)	-0.86±0.19 (0.11)	0.37±0.25 (0.23)
Mean	6.02±0.27	0.25±0.12 (0.04)	0.32±0.16 (0.06)	-0.06±0.22 (0.12)	0.31±0.18 (0.07)	-0.59±0.23 (0.12)	0.48±0.25 (0.24)

[1] Standard errors are from the refit and initial fit (in brackets)

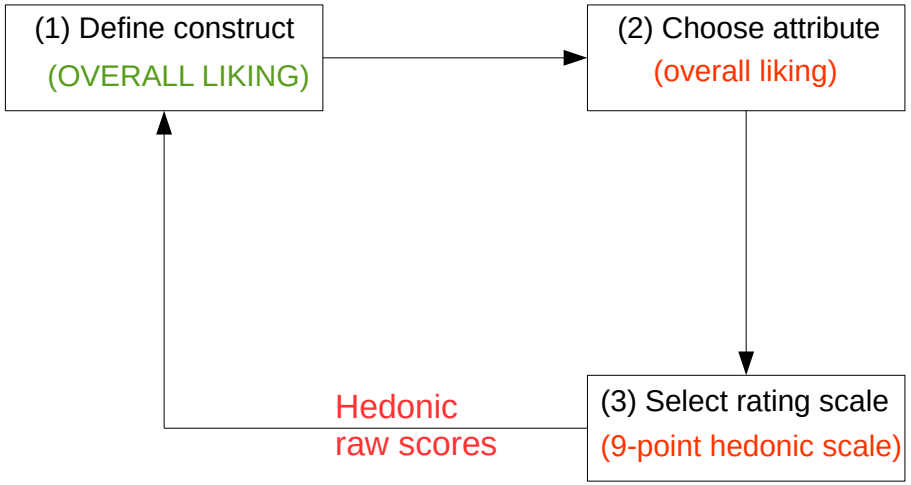

Fig. 1. Common approach of measuring overall liking

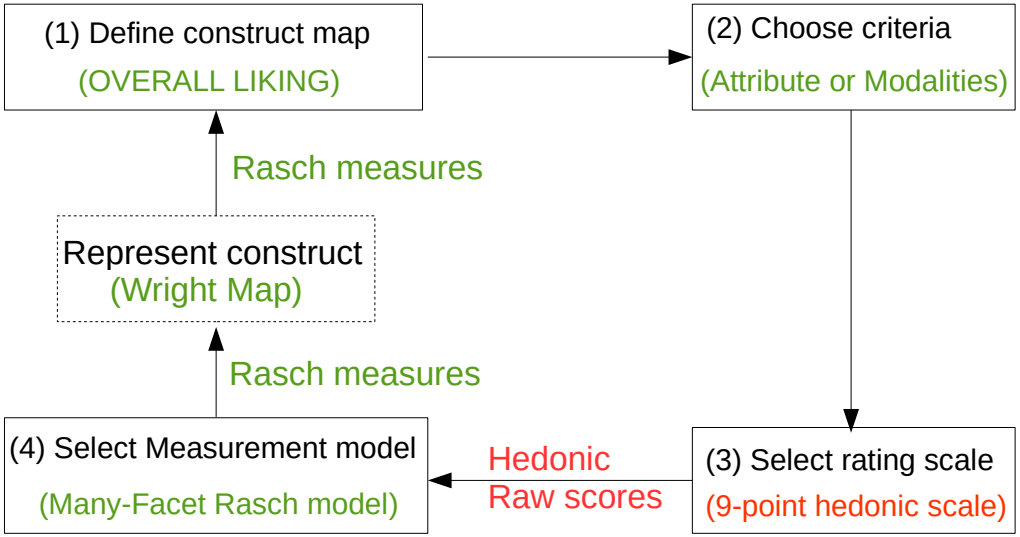

Fig. 2. Construct modelling approach for measuring Overall Liking

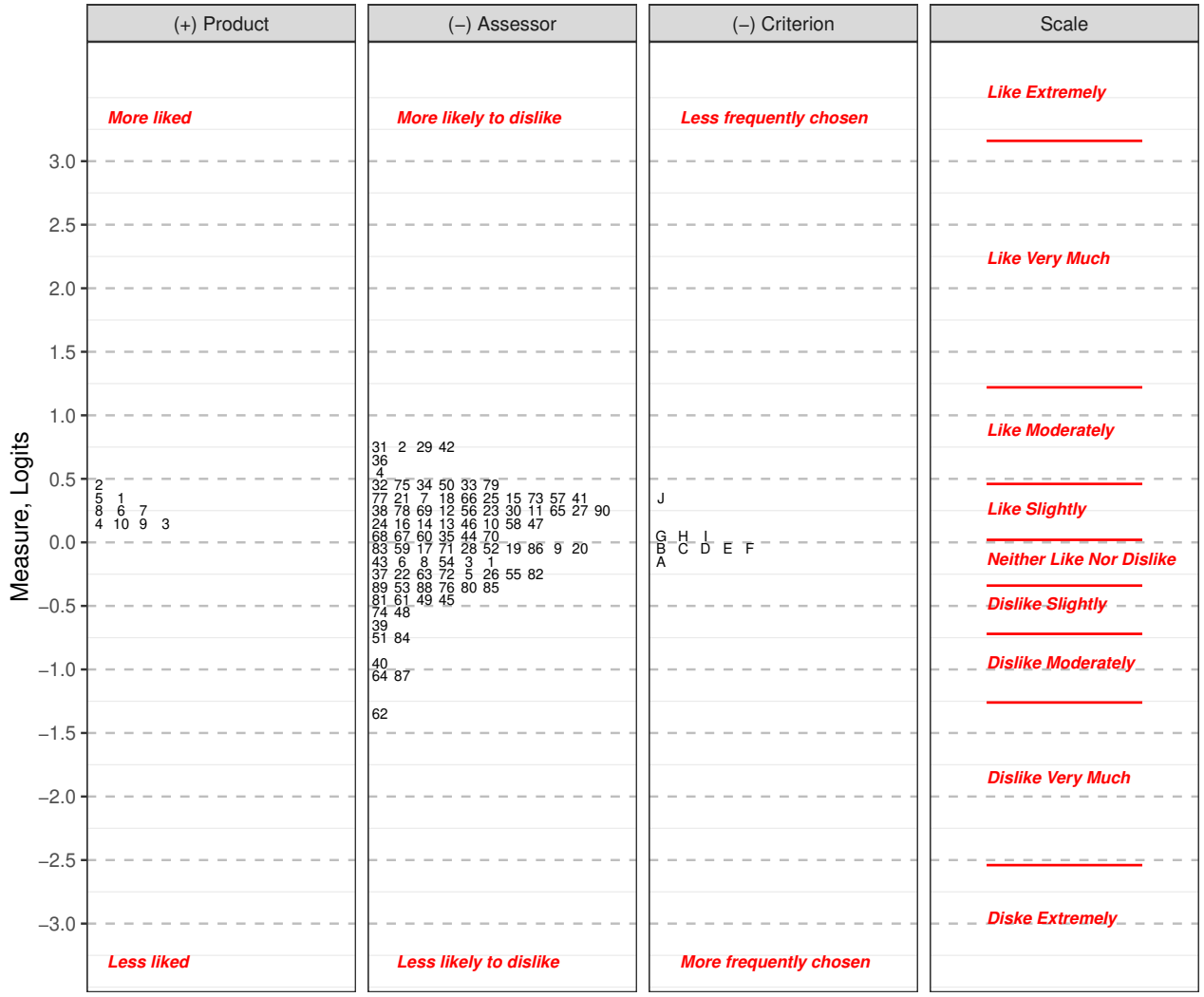

Fig. 3. Many-Facet Wright Map for IAM-R9. The three facets from left to right. Numbers 1-10 indicate products in Product facet and assessors 1-90 in Assessor Facet. Attributes are A: Hardness; B: Juiciness; C: Saltiness; D: Fibrousness; E: Redness; F: Typical Aroma; G: Aftertaste; H: Typical Flavour; I: Marbling; J: Sweetness. "Scale" represents the functioning of the scale, with horizontal lines representing Rasch-half-point thresholds where the average score is half way between two adjacent categories.

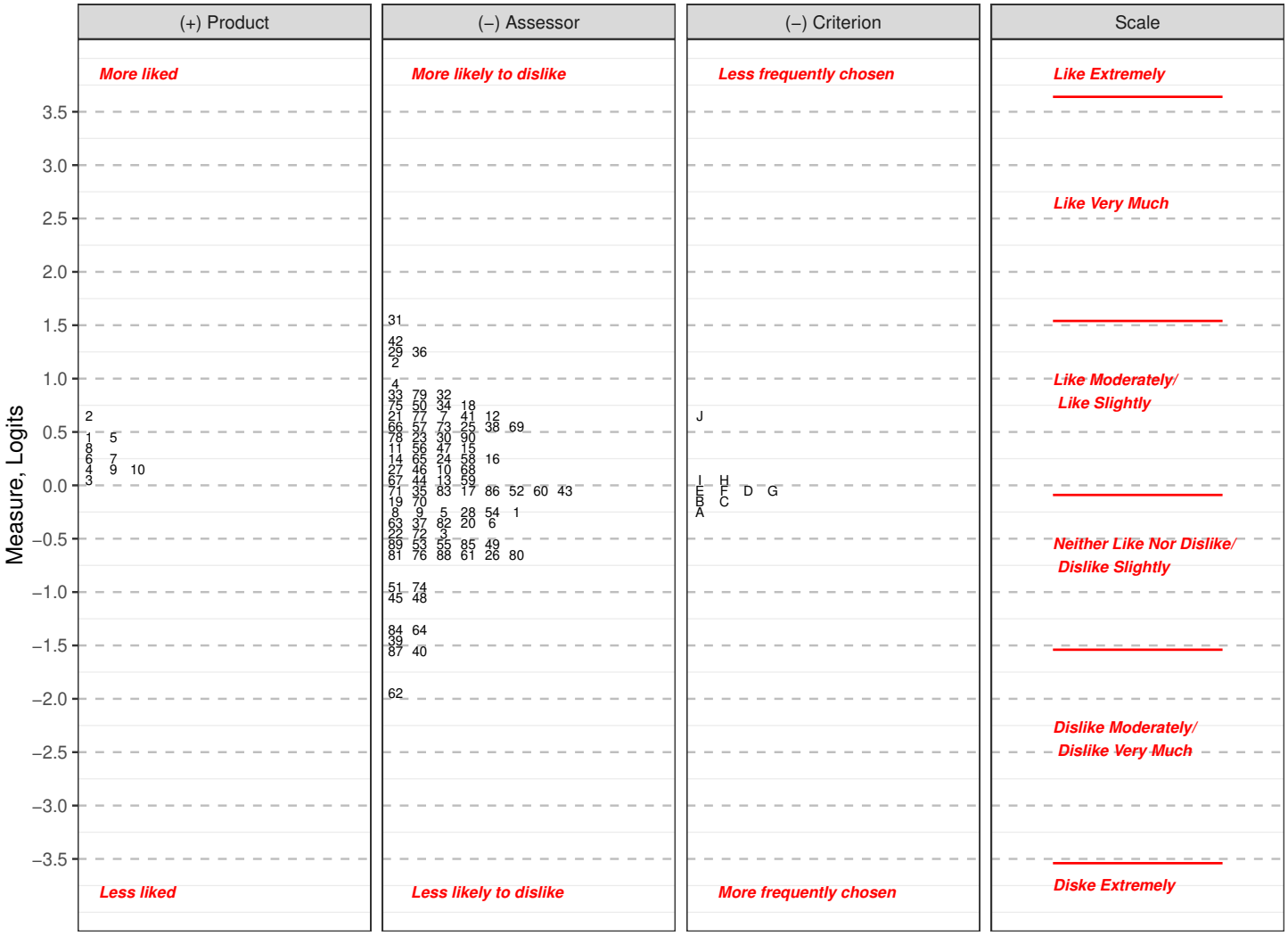

Fig. 4. Many-Facet Wright Map for IAM-R6. The three facets from left to right. Numbers 1-10 indicate products in Product facet and assessors 1-90 in Assessor Facet. Attributes are A: Hardness; B: Juiciness; C: Saltiness; D: Fibrousness; E: Redness; F: Typical Aroma; G: Aftertaste; H: Typical Flavour; I: Marbling; J: Sweetness. "Scale" represents the functioning of the scale, with horizontal lines representing Rasch-half-point thresholds where the average score is half way between two adjacent categories.