International Dairy Journal 97 (2019) 230–237

Contents lists available at ScienceDirect

International Dairy Journal

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

Application of the Rate-All-That-Apply (RATA) method to differentiate the visual appearance of milk powders using trained sensory panels

Rachel M. Traill*, Maree S. Luckman, Nicolene A. Fisk, M. Peng [1]

Fonterra Research and Development Centre, Fonterra Co-operative Group Limited, Private Bag, 11029, Palmerston North, New Zealand

ARTICLE INFO

Article history:
Received 28 February 2019
Received in revised form
27 May 2019
Accepted 27 May 2019
Available online 3 June 2019

ABSTRACT

The appearance of food products has a critical impact on consumer perception and acceptability; however, there is little information available that robustly measures the appearance of food from a human perspective with the potential to quantitatively inform product developers on how their products are visually perceived and directions for modification. This study demonstrates a successful case of trained sensory panels differentiating the visual appearance presented in 145 non-reconstituted milk powder samples, representing 11 different types of milk powders, using the Rate-All-That-Apply (RATA) sensory method. The results obtained from nonparametric statistical analyses implied significant differences in certain attributes, yielding a clear discrimination between the different milk powders. Furthermore, an additional study, consisting of four repeated assessments of 10 selected samples, suggested high test-retest reliability within and between the testing days. Overall, results from this study indicate the potential of RATA for characterisation of visual appearance of food products using trained panellists.

1. Introduction

Visual appearance of food products has a critical and perhaps immediate impact on consumers' acceptability and hedonic perception (Hutchings, 1977). Feedback from customers (who co-pack consumer milk powders) indicate milk powder is the type of product for which appearance is particularly important. Compared with food products that are consumed from a package, consumption of milk powders requires a much closer consumer–product interaction; the consumer purchases the powder and then utilises it within a home application such as drinking milk, tea, coffee, yoghurt or even cooking. Consumer home use tests indicate consumer's assessment of quality through optical perceptions of the product is thus inevitable and the appearance of the milk powder can give the first impressions of how the milk powder will reconstitute and therefore the level of quality. Variation in the appearance of milk powder from purchase to purchase can also lead to the consumer concluding a product could be unsafe or counterfeit. For these reasons, co-packers indicate it is important that the appearance of their milk powders must be consistent.

Direct communications with various co-packers show that certain markets prefer a certain appearance of milk powders. Milk powder manufacturers can make a wide range of different looking milk powders to meet their customers' preferences. The problem is that there is not a reliable method to assess the appearance of milk powders from a human perspective.

To date, most studies on powder functionality and quality have focused on analytical measures involving instrumentation that include particle size distribution, angle of repose, flowability, bulk density and particle density (Pugliese et al., 2017; Sharma, Jana, & Chavan, 2012). While being useful, sole dependence on instrumental measures can potentially pose a risk in misjudging the product's sensory quality, because they often have broader dynamic ranges than human perception (Carrasco & Siebert, 1999). The unparallelled instrumental–sensory relationship appears to be typically true with visual characteristics. For instance, Carrasco and Siebert (1999) presented such an example when testing the sensory–instrumental correlation on turbidity, in which the panellists' perception reached saturation prior to the upper limit of the instrumental measures. In this case, differences detected by the instruments become irrelevant to consumers' judgements of the products.

* Corresponding author. Tel.: +64 6 903 1073.
E-mail address: rachelm.traill@fonterra.com (R.M. Traill).

[1] Current address: Department of Food Science, University of Otago, PO Box 56, Dunedin, New Zealand.

https://doi.org/10.1016/j.idairyj.2019.05.013
0958-6946/

Food companies are particularly interested in testing discrepancies in powder appearance such as between manufacturing plants; inconsistency in the powder appearance could be seen by the consumers as quality variation. For the above reason, the most appropriate tool in assessing the visual appearance of milk powders are sensory panels. Since Baldwin (1977) identified the importance of incorporating evaluation of appearance into quality assessment of commercially manufactured milk powders, dairy industries have widely acknowledged the impact of powder appearance on consumers' acceptability. Baldwin (1977) had described a visual method based on photograph standards and a scoring system based on the degree of lumps (0–5 scale) and cohesiveness (A–F scale). This assessment was made immediately upon opening the tin or sachet of milk powder and so the Baldwin's (1977) method did well in measuring the first impressions of the milk powder. However, there is more to be understood about the appearance of this type of product, such as the size of the powder granules, the compactness of the granules, or even how the product moves within a vessel.

For identifying appearance discrepancies between powders, assembling a panel for complete sensory descriptive analysis (Lawless & Heymann, 1998) can be cost-intensive and time-consuming. Food companies are also continuously looking for more cost-effective ways to understand the sensory characteristics of their products with speed and least effort (Valentin, Chollet, Lelièvre, & Abdi, 2012). Rapid methods that can discriminate and describe the differences between samples are in demand. While Valentin et al. (2012) explained that rapid methods do not show small differences in sensory attribute intensities (as they would in descriptive methodologies), rapid methods do discriminate on larger differences. Valentin et al. (2012) also explained that rapid methods often do not have a training and calibration phase to ensure data quality from panellists; instead, larger panel numbers are used to give confidence in the data. With descriptive analysis smaller panel sizes are required ($\leq$20 sensory screened panellists) and sometimes is required to train and calibrate the panellists using reference samples to represent sensory attributes and their intensities (Stone, Bleibaum, & Thomas, 2012).

Over the last decade, researchers have developed some rapid profiling methods, and of these, Rate-All-That-Apply (RATA) is reported to be an effective method for providing sensory characterisation of products (Ares et al., 2014a). RATA is a variant of the Check-All-That-Apply (CATA) sensory characterisation method. In a CATA task, the respondent is presented with a list of sensory descriptive terms and asked to give a binary response (e.g., Yes or No) to the applicability of each term to the tested sample. A “No” response indicates when the attribute is not applicable, to reduce decision making complexity made by respondents (Jeager et al., 2014). Reinbach, Giacalone, Ribeiro, Bredie, and Frøst (2014), in awareness of limitations due to the dichotomised data, modified the CATA method by asking for an intensity rating on a 15-point intensity scale for the selected attributes. Later, Ares et al. (2014a) proposed the use of a three-point anchors on the intensity rating – low, medium and high, to retain the simplicity of this methodology. The three-point RATA, in comparison with CATA, was shown to yield more stable sample and term configuration, broader use of the attributes, and better sample discrimination (Ares et al., 2014a).

However, this previous study by Ares et al. (2014a) used 300 consumers. Many commercial companies do not have the capability of recruiting hundreds of people to conduct RATA testing to understand their products. Giacalone and Hedelund (2016) explored the RATA method with semi-trained assessors with promising results. This study extends the Giacalone and Hedelund (2016) findings further, with the complexity of understanding the appearance of milk powders.

The present study was designed to determine the applicability of RATA by trained sensory panellists, for assessing the visual appearance of 11 different milk powder types (a total of 145 non-reconstituted milk powder samples). The study comprised two facets that tested the method in terms of sample discrimination and reproducibility. RATA was chosen because this method would not require the development of reference standards to represent the scale of each attribute as required for descriptive analysis. However, unlike some other rapid methods, RATA does give a measure of which visual attributes are present in each sample and a basic idea of the intensity (Ares et al., 2014a). Since RATA is classified as a rapid method, it is understood this method would not give the level of detail of descriptive analysis (Valentin et al., 2012), but the method could still achieve a sufficient level of understanding of differentiation between milk powders and what characterised them as similar or different.

2. Materials and methods

2.1. Samples

This study tested a total of 145 milk powder samples from 11 different milk powder types. Eight to sixteen replicate milk powders (manufactured between October and December 2014) were selected from each milk powder type for testing.

The powders were stored in aluminium foil bags and subsampled (10 g) one-day prior to the testing into petri-dishes (90 mm diameter × 10 mm height) that were sealed with clear tape and then placed in the dark to protect from light until immediately before the testing. Each sample was labelled with a unique three-digit code.

2.2. Approach

Due to the large number of samples and RATA being an untested method for the intended purpose, the method was altered to suit the tools and resources of a commercial company; the approach to milk powder appearance using RATA is summarised in Fig. 1. The purpose of this was to train the testing sensory panel (referred to as “the panel”) on the visual extremes across all milk powder types, form a list of visual attributes applicable to the powders, and then understand how these visual attributes could be rated. The first two steps were to establish the panel and use expert assessors to screen 145 powders down to a training set for the panel. The panel was then used to generate the RATA descriptors from the training set of powder and were then trained and calibrated before assessing the 145 powder samples.

In addition to the main study, two sessions were conducted specifically designed to generate data for performance quality checking based on a set of 10 samples. These were used to determine the reproducibility and repeatability of the panel.

2.2.1. Panellists

A total of 33 trained sensory panellists participated in the current research. Panellists were aged between 25 and 73 years old (mean = 49.16), 26 were female, 7 were male. All panellists had been trained in descriptive analysis of flavours and tastes for a range of dairy products, but with no prior experience in either visual assessments or the RATA methodology. All panellists reported normal visual acuity.

Of the 33 panellists, 11 participated in the sessions of attribute-generation, which were held prior to the main studies. All 33 panellists participated in the RATA evaluation of powders, and of these 27 also participated in the repeatability and reproducibility

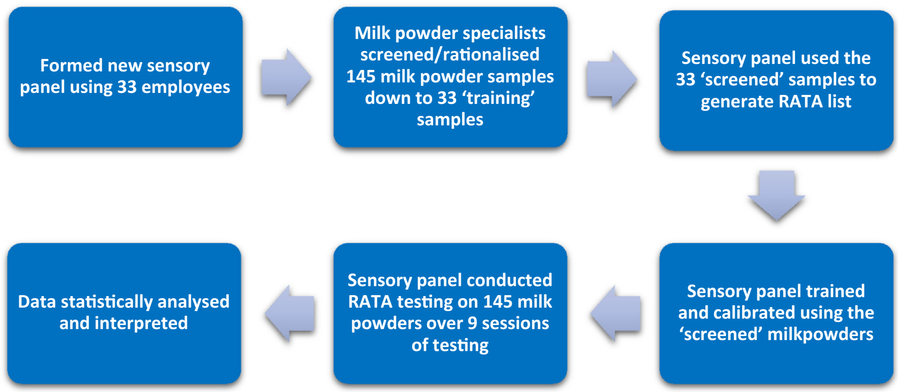

Fig. 1. Summary of milk powder appearance testing.

study; in each case a maximum of 14 panellists conducted assessments at one time due to the availability of sensory booths.

2.2.2. Sample screening

To provide a set of milk powder samples that would give panellists an understanding of the visual extremes they might experience of future sample assessments, a team of milk powder specialists ($n > 8$) were engaged. Eleven sessions were run in which all specialists were presented all the milk powders of one type (within a session). They were asked to sort the samples based on their visual similarities or differences (Chollet, Lelièvre, & Abdi, 2011) and then describe why they grouped the samples in a certain manner. The specialists were not told the type of milk powder they were viewing to eliminate any bias.

For each milk powder type, the data were collated and analysed using Multi-Dimensional Scaling (MDS) in XLSTAT (XLSTAT Version 2012.5.01, Add-insoft 1995–2012) to gain a consensus visualisation of sample similarities and differences. A minimum of 3 milk powders from each milk type (for example MP1, MP2, etc.) were selected to develop the attribute list for the trained panel. If the milk powder specialist did not find any differences between samples from within a milk type, 3 samples were randomly selected.

2.2.3. Attribute generation and training on RATA

The study began with two attribute-generation sessions. Each session took 1.5 h with the involvement of 11 panellists. A total of nine attributes were generated and defined and agreed upon by the panellists. In these two sessions, it was observed that the manner of manipulating the samples in the sealed petri dishes had a considerable effect on their resulting appearance. This prompted standardisation of the protocol of evaluating each attribute.

The protocol established comprised four steps: (i) static evaluation of the petri dish, (ii) evaluation during sideways shaking of the petri dish, (iii) static evaluation after sideways shaking and (iv) vertical positioning of the petri dish, a single gentle tap and then sliding the petri dish to move the powder away from the outer edge. Table 1 presents the list of attributes generated, along with their definitions, scale and stage of the regulated evaluation protocol.

While the powder colour was assessed by panellists, the colour data were not subsequently included in the statistical analysis. Colour was tested by instrumental analysis to mitigate non-reliable assessments between panellists due to any variations in lighting and shadowing of the individual booths. This information was not revealed to the panellists; the attribute remained included in the list so that colour did not become a psychological barrier for the panellist to measure the powder differences.

A 1.5 h training was then administered to all panellists across different sessions. The training entailed explanation of each attribute generated by the original panel, and familiarisation of the protocol of assessing each attribute. In addition, the panellists were given a set of samples and asked to conduct a ranking task for each of the attributes. The aim of the ranking exercise was to prompt understanding of the definition of the attribute, rather than standardising the panellists' perception of each intensity category. For the RATA task, the panellists were instructed to use the intensity level according to their own definition.

2.3. RATA evaluation of powders

Nine one-hour sessions were undertaken each comprising RATA evaluation of 20 milk powders. These milk powders consisted of 17 randomly selected samples from the sample set, one warm-up sample (which was always presented first), one randomly-selected duplicate sample, and one control. The warm-up sample was the same across all the panellists. Data obtained for the warm-up samples were omitted in the subsequent analyses. The duplicated sample was randomly selected among the 17 milk powders; it served as an assessor for the panellists' consistency in performance within a session. In addition, a fixed control sample (MP4 D1437-1) was included in all sessions to monitor the level of consistency in the panellists' performance across (between) sessions.

Apart from the warm-up sample, the remaining 19 samples were presented sequentially to the panellists in a randomised order. The randomisation across the panellists was designed according to the William's Latin Square, so that any carry-over or sequential effects were mitigated. A 10 min break was enforced after the 10th sample.

For each sample, the panellists were asked to respond "yes" to the attribute(s) they considered appropriate for describing the sample and then to rate the intensity of the attribute(s), using a 3-point scale ('low', 'medium', and 'high') (Table 1). The order in which the terms were presented was randomised for each product and participant, as suggested by Ares et al. (2014b).

The response data were collected by an internet-based software – SurveyMonkey®. The evaluations were conducted in standard, individual sensory booths under artificial daylight illumination that gave adequate viewing of the samples; however, as an additional precaution against possible discrepancies in the lighting condition, each panellist was asked to attend the same designated booth for all the testing sessions.

2.4. Establishing panel number and performance

The RATA method has been thought of as useful for consumer evaluation of samples. This study looks at the possibility of using trained panellists. There is an assumption there would be less

Table 1
Attribute descriptors, definitions, evaluation protocol and scale.[a]

Attribute	Definition	Scale definition - extremes	
		1	3
Colour	The shade of colour of the powder (static)	Light	Dark
Particle size	The perception of the powder particles from fine particles to coarse/large particles (static)	Fine particles	Coarse particles
Flowability	The easiness of the powder moving in the petri-dish (shaken)	Low free moving	High free moving
Clumpiness	The tendency to form clumps (shaken/static)	Low amount of clumping	High amount of clumping
Clumps different sizes	The tendency to form clumps of different sizes (shaken/static)	Less different	More different
Density	The compactness the powder (shaken/static)	Airy	Dense
Cracked surface	The quantity of cracks on the powder surface (shaken/static)	Low amount of visual cracking	High amount of visual cracking
Plate coverage	The coverage of the base of the petri-dish (shaken/static)	Low amount of coverage	High amount of coverage
Complete edge	The shape of the outer edge (slide)	Fractured edge	Complete edge

[a] Evaluation protocol is given in parenthesis after the definition: static, static position; shaken, during the process of being shaken sideways: shaken/static, static position after being shaken sideways; slide, hold the Petri dish vertically and gently tap it once, then slide the dish to move the powder away from the outer edge. Note on scale definition: 0 was utilised when a panellist deemed an attribute not relevant to a sample.

variability in trained panel evaluations (individually and as a group) than consumers, and therefore less people required to test. A power calculation was conducted based on the size of panel that could be practically utilised within the available sensory facility and an assumed variance of 10% of the range (0–3). The aim of the study was to identify differences between powder types that consumers would react to, so an absolute difference between powder means of 0.5 was considered the minimum difference of importance. The level of significance and power set at 0.05 and 0.90 respectively was deemed sufficient for this study. The actual power for the final test of powder type differences based on 145 samples is also reported (see Supplementary material, Tables S3 and S4).

To understand if the RATA testing method was repeatable (panellist can repeat their results within a session) and reproducible (the panel result can be reproduced between sessions), the same 10 selected samples were assessed four times by each panellist, being duplicate assessments (d1 and d2) of each sample in two sessions (Sessions J and K).

The 10 samples were randomly selected from the 11 different milk powder types. To ensure the performance measures were valid for all powder types, there was one sample from all powder types, except for MP7.

The data collection for this was procedurally identical to that described for RATA evaluation of powders.

2.5. Data analysis

In these data, the absence of the attribute is scored as zero (0) and intensity for attributes that are present are scored as 1, 2 or 3.

After data and panel performance checks were completed, overall means, per sample, per attribute were calculated. Thus, the resultant means were used as a continuous variable derived from categorical data. Vidal, Ares, Hedderley, Meyners, and Jaeger (2018) stated that “The comparison of RATA data using mean scores and Dravnieks' scores showed no advantage of the latter and it is recommended that simple mean scores are used”. That is, for each sample ($n = 145$) the mean intensity was calculated using score data collected on the 0–3 scale of all evaluations from the 9 main evaluation sessions. Each sample mean was calculated from between $n = 25$ and 62 assessments, some being the mean of duplicate assessments. These means were used in the analysis of powder types with $n = 8–15$ sample means within each powder type.

A generalised linear model (GLM) using Minitab Release 17.2.1 (Minitab Inc, State College, PA, USA, 2015) was used to test for differences between powder types. The error estimate was therefore the pooled between sample within milk powder type variance. The standardised residuals from the GLM were checked that they met the requirements for the assumptions of GLM and were sufficient to give confidence in the use of the method for these data.

For those attributes that showed significance through GLM, a comparison between powder type means to identify which types were significantly different was conducted using Tukey's test, controlling for a 5% simultaneous error rate (Rafter, Abell, & Brasselton, 2002; Tukey, 1949). The outcomes can be found in Supplementary material, Table S1.

Principal component analysis (PCA) (Manly, 1994) is a dimension reduction method used to generate a “map” allowing visual comparison of multi-dimensional data in fewer dimensions, usually two. A map summarising all of data from the 8 powder appearance attributes measured (colour was excluded) for all the milk powders samples was generated. This enabled an understanding of which powder appearance attributes define the differences/similarities between the milk powders and the amount of variability between samples from within, and between, milk powder types.

The data collected in the two sessions conducted specifically to assess panel performance were analysed at three different levels.

Overall panel alignment was assessed using XLSTAT Multiple Factor Analysis (MFA) (Abdi, Williams, & Valentin, 2013; Robert & Escoufier, 1976). MP7 was not used in this analysis because it had 8 samples whereas other powder types had 13 or 15. As the powder type means were being used it was determined the precision around the different powder types would be different, therefore MP7 was not included in the MFA. The panellist mean scores for each attribute by powder type was calculated and the resultant MFA plot and metrics assessed. Specifically, RV coefficients were checked to identify any panellists who were not aligned to the overall panel trends; RV values are a form of multivariate correlation coefficients (Abdi, 2007; Robert & Escoufier, 1976).

Panellist reproducibility was checked to determine ‘between sessions’ variation as an important metric of impact on the comparison between samples tested in different sessions. To do this, the ‘within’ session attribute mean, per panellist, per sample was calculated and a comparison made of these session means. Reproducibility was assessed by a Pearson's correlation between the means from sessions J and K. A high correlation coefficient would indicate that overall the panellists were achieving a similar result on both days.

Additionally, a tally was made of exact matches between mean scores. The number of times a panellist mean score were an exact match, greater than ± 0.5 and greater than ± 1 between sessions was counted and expressed as a percentage of all assessments by a panellist (Table 3). To meet the aims of this study it was deemed that if the difference between mean scores between days was within ±0.5 that was sufficient to indicate alignment.

Repeatability is a measure of the ability of a panellist to repeatedly give the same score on the same sample. A strong repeatability value indicates the panellists can consistently use the scale to measure the parameter and is therefore one important component of assessing data quality. Repeatability was assessed at an individual panellist level by analysis of scores for duplicate pairs within sessions J and K. On each powder appearance attribute, each panellist made 20 duplicate pair assessments (2 sessions, 10 samples).

Exact match between the duplicate milk powders for an attribute indicates a panellist gave the same score for the pair of duplicates within a session, whereas a ±1 match means that for a sample duplicate pair scores were within 1 of each other. It was determined that, for this study, acceptable repeatability performance was defined as a panellist achieving 90% of their duplicate assessments within ±1 on the rating scale across all attributes. The tallies across all attributes are shown in Supplementary material, Table S2.

To better understand what attributes were proving most difficult for panellists to assess, an analysis of the performance in different attributes were undertaken. Cross tabulation and tallies by attribute and panellist were conducted, but were not considered for acceptance of the panellist as performing sufficiently for data to be retained.

3. Results and discussion

3.1. Powder appearance attribute differences between milk powder types

Differences were found between powder types for all 8 attributes measured (Fig. 2A–H), the attributes being: particle size, clumpiness, clumps different sizes, flowability, density, complete edge, cracked surface and plate cover. While the definitions of these attributes are specific, the results indicate that plate cover and flowability are highly correlated, as well as density and clumpiness. This indicates a powder which has high flowability will have a high level of plate cover; and a powder that is viewed as being clumpy is also viewed as being dense.

On average, milk powder types MP11, MP5 and MP7 had the highest amount of plate cover (Fig. 2H; 2.93, 2.64 and 2.61, respectively) and flowability (Fig. 2D; 2.84, 2.55 and 2.55, respectively), where MP10 had the least amount of plate cover and flowability (1.15 and 1.10, respectively).

With respect to density (Fig. 2E), milk powder types MP11 and MP7 were deemed airy (1.16 and 1.39, respectively), whereas the MP10 was deemed to be dense (2.86).

Of all the milk powder types, MP11 and MP7 were the least clumpy (Fig. 2B; 0.80 and 1.13, respectively) and MP10 was the clumpiest (2.75). MP11 and MP7 also had the least clumps different sizes (Fig. 2C; 0.78 and 0.98, respectively). Whereas, MP10 and MP6 showed the most clumps different sizes (2.19 and 1.96, respectively).

Milk powder types MP10, MP6 and MP4 had the highest level of complete edge (Fig. 2F; 2.54, 2.32 and 2.33, respectively), whereas MP11 (mean = 0.98) had the least. MP11 also had the lowest amount of cracked surface (Fig. 2G; 1.17), whereas MP10 and MP4 had the highest (1.82 and 1.73, respectively).

Small differences were found for particle size (Fig. 2A), ranging from 1.58 for MP11 to 2.04 for MP10. Overall, therefore, particle size had the smallest impact in differentiating between milk powder types.

Table 3
Correlation between the two sessions for sample/panellist means.

Attribute	Correlation (n = 270)	Match (%)		
		Exact	±0.5	±1.0
Particle size	0.75	74.4	90.0	98.9
Clumpiness	0.69	46.2	80.9	95.2
Clumps different sizes	0.72	50.0	75.0	93.7
Flowability	0.80	51.0	81.5	98.1
Density	0.78	49.3	83.3	98.2
Complete edge	0.69	37.8	72.6	90.7
Cracked surface	0.64	44.0	77.0	90.0
Plate cover	0.77	43.0	83.0	83.0

3.2. Milk powder type differentiation

Overall, the RATA method was shown to differentiate gross differences between samples for powder appearance. A PCA, using the raw 145 sample panel means, was conducted to summarise the 8-attribute multivariate data. Of the variance in the data, 94.3% was explained in the first component (PC1) and 2.1% in the second component (PC2). By summarising all the powder appearance data in a PCA (Fig. 3), 3 clear exemplars of powder appearance, relative to each other, were apparent. Exemplar 1 is a fine powder, free flowing with no clumps, Exemplar 2 presents a small number of clumps, but still slightly free flowing and Exemplar 3 powder has dense soft clumps of varying sizes.

In general terms, most of the milk powder types look similar to Exemplar 2. Most of MP11 milk powders are Exemplar 1, and MP10 milk powders are Exemplar 3. Of all the data generated, MP8 is the typical Exemplar 2 powder, and on the spectrum of milk powder types, it is seen to be midpoint between MP11 and MP10. The powder appearance attribute mean intensities for the three key milk powders types are summarised in Table 2.

3.3. Panel size and performance

3.3.1. Panel size

Power calculations showed that a panel of 33, assuming a standard deviation of 0.3, had power of 90% to detect differences between 11 sample means of at least 0.34. This was within the size of difference thought to be of importance for this project. There was very small loss of power (86.5% compared with 90%) with the eventual sample size of 31 trained panellists.

To determine overall panel consistency, the outputs from an MFA were assessed (Fig. 4). A review of the RV coefficients showed one panellist, RTER, was not correlated to most other

Table 2
Powder appearance attribute mean intensities for the three key milk powder exemplars.[a]

Powder exemplar	Milk powder type	Particle size	Clump	Clumps different sizes	Flow-ability	Density	Complete edge	Cracked surface	Plate cover
Exemplar 1	MP11	*1.58*	*0.80*	*0.78*	**2.84**	*1.16*	*0.98*	*1.17*	**2.93**
Exemplar 2	MP8	**1.97**	1.81	1.59	2.04	1.89	1.99	1.53	2.18
Exemplar 3	MP10	**2.04**	**2.75**	**2.19**	*1.10*	**2.86**	**2.54**	**1.82**	*1.15*

[a] Values in bold and italic indicate fonts indicate highest and lowest (or not different from highest and lowest) mean scores, respectively, compared with all 11 milk powder types.

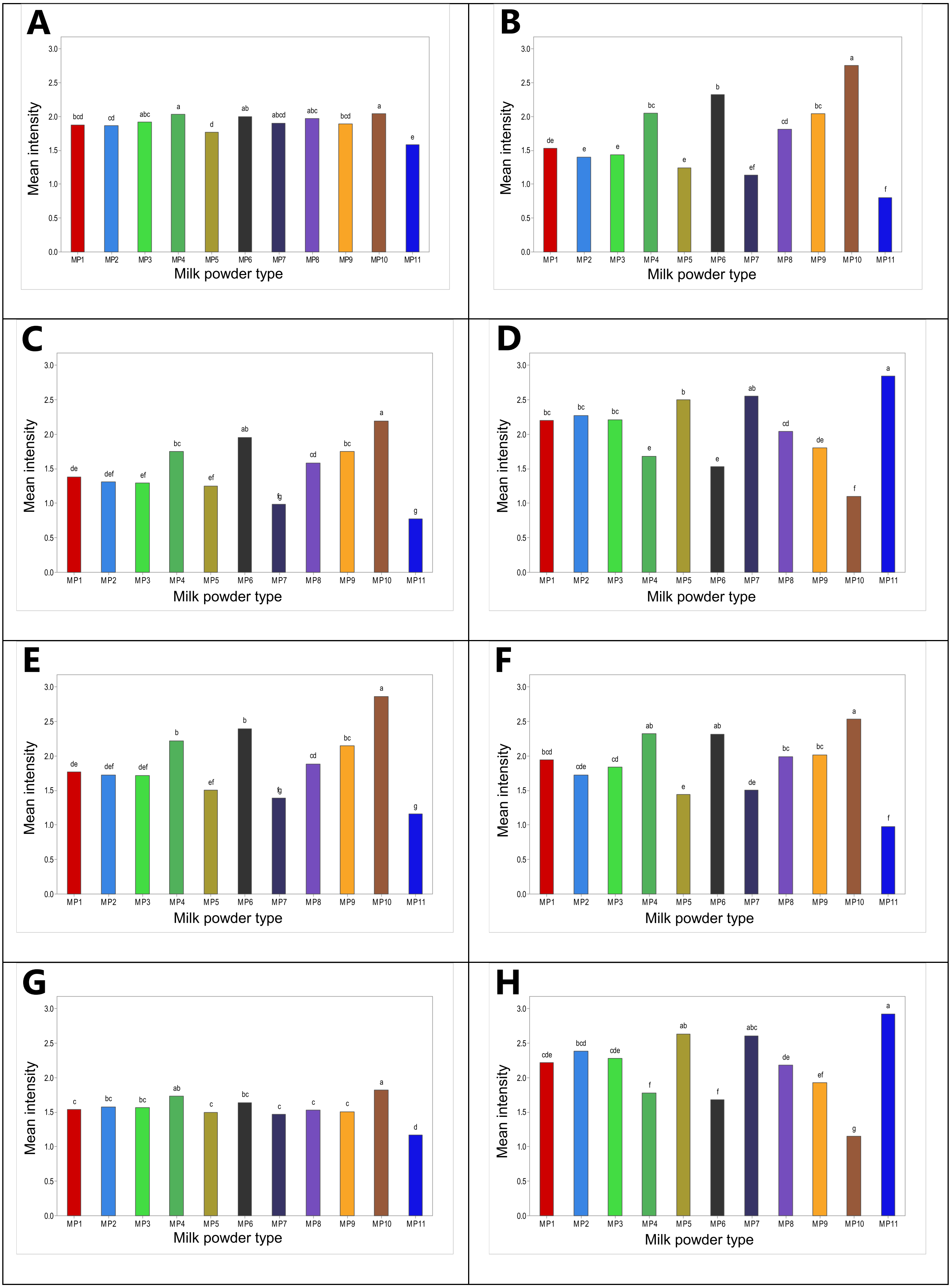

Fig. 2. Powder appearance differences between milk powder types: A, particle size; B, clumpiness; C, clumps different sizes; D, flowability; E, density; F, complete edge; G, cracked surface; H, plate cover. Values are the panel means for each milk powder type tested with the letters above each bar on the charts representing the Tukey's mean separation; means that do not share the same letter are statistically significantly different ($p < 0.05$ family error rate).

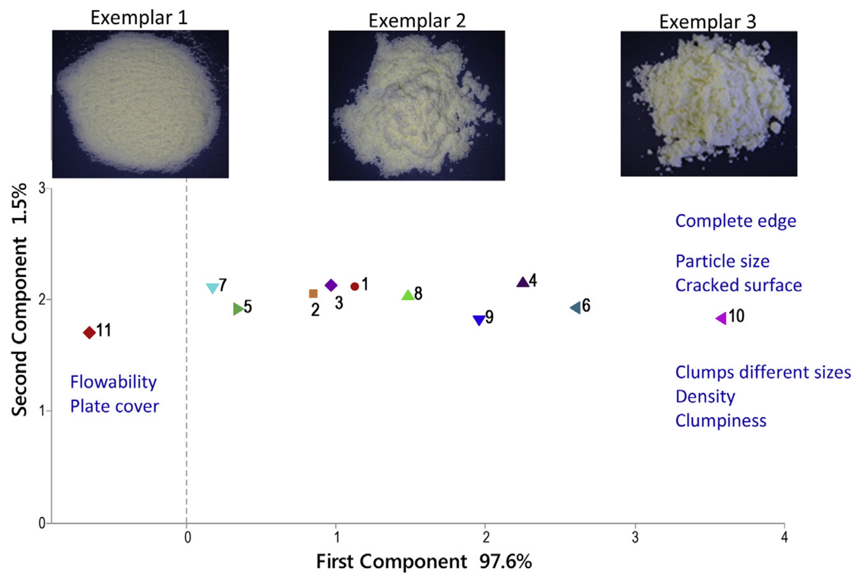

Fig. 3. Milk powder appearance principal components analysis showing the three exemplar powders.

panellists, so was deemed to be non-aligned; all data from this panellist were removed for the remaining data quality checks and from the main analysis of milk powder type comparison analyses.

While panellist RTER was consistent in the assessment of the milk powders, the multi-factor analysis (MFA; Fig. 4) shows that RTER was assessing the samples differently from the rest of the panel, as seen by the distance from the others in the important (71.7%) Factor 1 direction (x-axis). RTER overall RV was 0.712, whereas all other panellists RV overall ranged from 0.830 to 0.982. This outcome is seen due to the panellist defining the attribute or its scale differently from the rest of the panel. For this reason, RTER's data were excluded from the calculations and analyses of the milk powder type and sample means.

3.3.2. Between sessions - reproducibility

A coefficient of less than 0.7 for between session correlation was determined to indicate moderate variation in the data between sessions J and K. It was seen (Table 3) that for clumpiness, complete edge and cracked surface the correlation coefficient was below this level; however, similar differences were still found between the milk powders. The mean scores for sample*panellist between testing days for particle size, clumps different sizes, flowability, density and plate cover were strongly correlated.

The alignments of between-session mean results per panellist per sample are shown (Table 3) as the % match at 3 levels of attainment. Of the results of one panellist (JBDE), 13.3% were outside the upper reproducibility limit, therefore JBDE's results were also excluded in the overall final analyses.

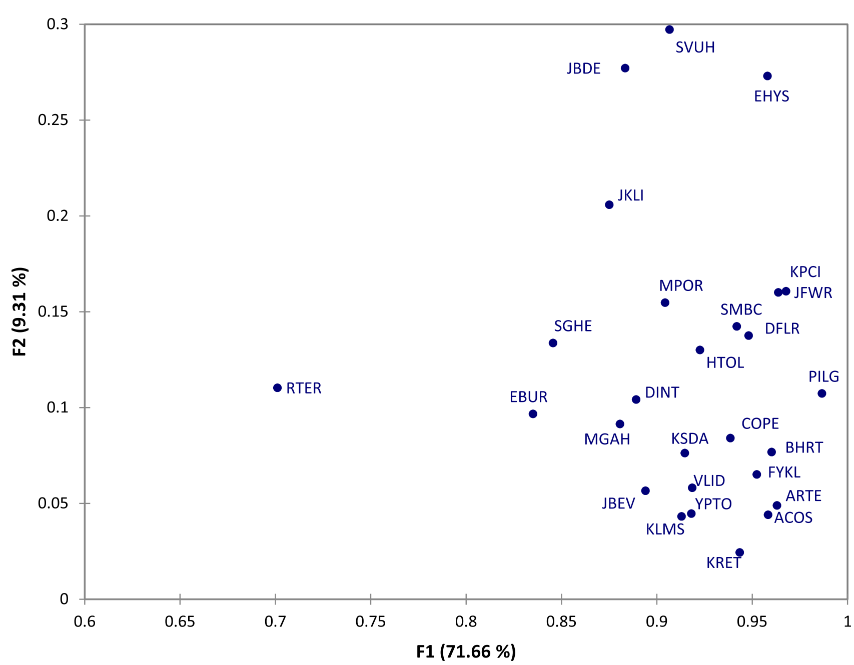

Fig. 4. Multiple factor analysis of panellists by sample means for 9 powder appearance attributes.

3.3.3. *Repeatability*

The ability for panellists to assess samples in the same way, within the same session, indicates repeatable results. It is the first level of achieving reliable assessments from the research. The measure of acceptance for repeatable assessments was set at a minimum of 90% of the duplicate pairs were within ±1 of each other. Across all nine attributes, all except one panellist achieved over 90% of their results within this limit. Panellist KPCI achieved 89.4% within the ±1 limit, which was considered substantially achieved, therefore KPCI's data were retained.

In general panellists performed well with all except of three panellists achieving the target performance for at least 7 out of 9 attributes. Clumps different sizes and complete edge were the most difficult for panellists to assess, with 19 and 20 (respectively) of the 28 panellists achieving the repeatability measure, but for flowability and density all panellists met the measure.

4. Conclusions

The RATA method was a robust method that described and measured the appearance of a range of milk powder types and differentiated between them. The method showed differences between milk powder types with confidence. Overall it can be seen that the differences between milk powders were clearly defined. The implication is that the RATA method could define visual differences between milk powders effectively.

To use the RATA method with a small number of panellists, it is necessary for the panellists to view the extremities of the samples first, then can both generate the required attributes and understand what is considered high and low in its levels of intensities. While this activity took some time, it is considerably less time and resource than is required for descriptive analysis. A further step in assessing the RATA method would be to use sensory panels that have no experience with the method or product to see if there is a similar outcome; this could help determine if a product to be assessed truly needs to be introduced to the testing panel first to obtain reliable results.

Overall this study has led to the development of a quality control method for the appearance of milk powders. The differentiation found within this study helped identify the samples used to develop photographic standards. These photographic standards, along with a method for sample preparation and assessment was further developed and distributed to milk powder quality laboratories. Technicians within these quality laboratories were trained and calibrated on the preparation and assessment of milk powder against these photographic standards. This development of a quality method, means milk powder producers have a better understanding of what their milk powders look like and confidently supply milk powders with more consistent appearance to co-packers.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idairyj.2019.05.013.

References

Abdi, H. (2007). RV coefficient and congruence coefficient. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 849–853). Thousand Oaks, CA, USA: Sage.

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *WIREs Computational Statistics, 5*, 149–179.

Ares, G., Bruzzone, F., Vidal, L., Cadena, R. S., Giménez, A., Pineau, B., et al. (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.

Ares, G., Etchemendy, E., Antúnez, L., Vidal, L., Giménez, A., & Jaeger, S. R. (2014). Visual attention by consumers to check-all-that-apply questions: Insights to support methodology development. *Food Quality and Preference, 32*, 210–220.

Baldwin, A. J. (1977). Appearance of whole milk powder as related to physical properties. *New Zealand Journal of Dairy Science and Technology, 12*, 201–202.

Carrasco, A., & Siebert, K. J. (1999). Human visual perception of haze and relationships with instrumental measurements of turbidity. Thresholds, magnitude estimation and sensory description analysis of haze in model systems. *Food Quality and Preference, 10*, 421–436.

Chollet, S., Lelièvre, M., & Abdi, H. (2011). Sort and beer: Everything you wanted to know about the sorting task but did not dare to ask. *Food Quality and Preference, 22*, 205–520.

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.

Hutchings, J. B. (1977). The importance of visual appearance of foods to the food processor and the consumer. *Journal of Food Quality, 1*, 267–278.

Jaeger, S. R., Cadena, R. S., Torres-Moreno, M., Antúnez, L., Vidal, L., Giménez, A., et al. (2014). Comparison of check-all-that-apply and forced-choice Yes/No question formats for sensory characterization. *Food Quality and Preference, 35*, 32–40.

Lawless, H. T., & Heymann, H. (1998). *Sensory evaluation of food. Principles and practices* (pp. 227–246). New York, NY, USA: Chapman & Hall.

Manly, B. J. F. (1994). *Multivariate statistical methods: A primer* (pp. 76–81). Portland, MD, USA: Chapman & Hall.

Pugliese, A., Cabassi, G., Chiavaro, E., Paciulli, M., Carini, E., & Mucchetti, G. (2017). Physical characterization of whole and skim dried milk powders. *Journal Food Science Technology, 54*, 3433–3442.

Rafter, J. A., Abell, M. L., & Brasselton, J. P. (2002). Multiple comparison methods for means. *Society for Industrial and Applied Mathematics, 44*, 259–278.

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.

Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Applied Statistics, 25*, 257–265.

Sharma, A., Jana, A. H., & Chavan, R. S. (2012). Functionality of milk powders and milk-based powders of end use applications - a review. *Comprehensive Reviews in Food Science and Food Safety, 11*, 518–528.

Stone, H., Bleibaum, R. N., & Thomas, H. A. (2012). *Sensory evaluation practices* (pp. 250–260). New York, NY, USA: Academic Press.

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*, 99–114.

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: A review of new descriptive methods in food science. *International Journal of Food Science and Technology, 47*, 1563–1578.

Vidal, L., Ares, G., Hedderley, D., 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.