

COMPARISON OF DIFFERENT SENSORY DISCRIMINATION TESTING METHODS USED IN FOOD INDUSTRY

An application of triangle, tetrad, and duo-trio tests using various food samples



Master's thesis

Hämeenlinna, Bioeconomy and Business Development

Autumn 2021

Julie Marie Oca Rodriguez

Biotalous liiketoiminnan kehittäminen

Tiivistelmä

Tekijä Julie Marie Oca Rodriguez

Vuosi 2021

Työn nimi Erilaisten aistinvaraisen erotustestimenetelmien vertailu elintarviketeollisuudessa: kolmi-, neliö- ja pari-kolmitestien soveltuvuus erilaisten elintarvikenäytteiden arviointiin

Ohjaajat Tuija Pirttijärvi, Paula Koivisto, Tiina Hämäläinen

Elintarviketeollisuudessa tuotteita kehitetään jatkuvasti kuluttajien toiveiden ja vaatimusten sekä trendien mukaisesti. Muutoksia voi tapahtua tuotantoprosessissa, reseptissä, valmistusaineissa, pakkauksessa tai säilytyksen olosuhteissa. Nämä muutokset voivat aiheuttaa ei-toivottuja muutoksia tuotteen aistinvaraisiin ominaisuuksiin. On tärkeä, että tuotteen aistittava laatu pysyy hyvänä ja muutos ei aiheuta riskejä. Muutoksen vaikutus tuotteiden aistittavaan laatuun varmistetaan erotustestillä: kehitettyä tuotetta verrataan nykyiseen tuotteeseen ja selvitetään aiheuttaako muutos aistinvaraisesti havaittavan eron.

Kolmitesti on yksi yleisimmistä erotustesteistä. Kolmitestiä käytetään näytteiden pienten erojen tutkimisessa laajasti sekä Valion tuotekehityksessä että tuotannossa. Tuloksia käytetään päätöksenteon apuna laadunvarmistuksessa ja tuotekehityksessä. Testin järjestäminen on vaativaa, aikavievää ja tarvittava näytemäärä on suhteellisen suuri. Erotustestejä ja erityisesti kolmitestiä on tutkittu paljon viime vuosina. Kolmitestin luotettavuus on saanut osakseen kritiikkiä. Muita erotustestimenetelmiä on tutkittu ja kehitetty. Näistä syistä aistinvaraisen menetelmän valintaa yrityksessä oli syytä tarkastella uudelleen.

Tämän tutkimuksen tavoitteena oli testata ja verrata erilaisia elintarvikeyrityksessä käytössä olevia erotustestejä. Kolmen erotustestimenetelmän soveltuvuutta yleisten erojen tutkimiseen verrattiin Valion tuotteiden arvioinneissa. Tulosten perusteella kaikki kolme menetelmää soveltuivat maito-, juusto- ja maustettujen jogurtinäytteiden erojen tutkimiseen erotustestillä. Pari-kolmitestillä saatiin yhtä luotettavia tuloksia kuin kolmitestillä, vaikka tarvittava näytemäärä oli pienempi. Neliötestillä saatiin yhdenmukaisia tuloksia kolmitestiin verrattuna, joskin testissä tarvitaan enemmän näytettä. Arvioijien mukaan, neljäs näyte antoi lisävahvistuksen siitä, mitkä näytteet ovat erilaiset kuin kaksi muuta näytettä. Tämä tutkimus on tarjonnut sisäiselle raadille lisää tietoa erilaisista erotustestivaihtoehdoista ja harjoitussarjoja eri menetelmien käytöstä. Näitä kolmea menetelmää on hyvä hyödyntää erilaisissa tilanteissa jatkossa, jolloin sisäinen raati harjaantuu paremmin erojen tunnistamisessa eikä rutinoidu vain kolmitestiin.

Avainsanat aistinvarainen arviointi, erotustesti, kolmitesti, neliötesti, pari-kolmitesti

Sivut 70 sivua ja liitteitä 18 sivua

In food industry, products undergo different stages of development to meet the constantly changing trends and consumer demands. Modifications can be in the product's processing, recipe, ingredients, packaging, or storage conditions. Implementing these modifications can create unwanted changes in the sensory characteristics of the product. It is important to make sure that the current sensory quality of the product will be maintained and that a certain change will not be a risk. To verify whether the changes affect the product's sensory quality, discrimination testing can be implemented by comparing the modified product to the current version of the product available in the market.

The triangle test is one of the most common discrimination testing method widely used and it has been the main method used for sensory discrimination testing of food samples in Valio's product development and production plants. Results of triangle tests are used in the company to support decision making in quality assurance and product development. Organizing the triangle test is challenging, time-consuming, and the total amount of samples needed is relatively large. Triangle test has received mixed reviews from other experts being prone to many errors. With new methods being studied and developed, other alternative methods for discrimination testing must be reviewed.

This study's aim was to test and compare different discrimination testing methods used in food industry. The applicability of three overall difference testing methods triangle, tetrad, and duo-trio was compared using the company's own products. Results showed that the three methods can be used alternatively when the food type to be evaluated is milk, cheese, or mixed-flavored yogurt. Compared to triangle test, duo-trio test provided equally significant and reliable results, while only lesser number of samples were required. Tetrad test also provided equally significant and reliable results, while providing the assessors more confidence in their answers with the presence of the fourth sample as a confirmation. This study served as a useful learning experience for the internal panel, providing series of practice on performing different discrimination testing methods themselves. Alternatively using these three methods is a good practice to consider in the future, as the assessors start being accustomed to the routinely used triangle test. Performing different discrimination tasks will help improve the panel's performance in distinguishing sensory differences.

Keywords sensory evaluation, discrimination test, triangle test, tetrad test, duo-trio test

Pages 70 pages and appendices 18 pages

Table of Contents

1	Introduction.....	1
2	Aim of the study	3
3	Sensory science	4
3.1	Sensory evaluation.....	5
3.2	Sensory evaluation methods.....	6
4	Discrimination testing methods	11
4.1	Triangle test and its pitfalls	18
4.2	Methods of analysis	23
5	Materials and methods	26
5.1	Methods.....	27
5.2	Products	29
5.3	Recruitment	31
5.4	Participants	31
5.5	Location.....	33
5.6	Samples	33
5.7	Test questionnaire	37
5.8	Data collection	39
5.9	Data analysis	40
6	Results and discussion.....	42
6.1	Milk.....	42
6.2	Juice.....	44
6.3	Yogurt.....	46
6.4	Plant-based yogurt alternative (<i>gurt</i>)	50
6.5	Cheese.....	52
6.6	Summary	54
7	Conclusion and recommendations.....	59
8	Acknowledgements	63
	References.....	64

Table of Figures

Figure 1. Sensory science as a “link” to other fields of science	4
Figure 2. Main methods of sensory evaluation	6
Figure 3. Method selection based on research objective	6
Figure 4. Product-oriented tests vs. consumer-oriented tests	7
Figure 5. Seven-point facial hedonic scale	8
Figure 6. A sensory profile of two milk samples.....	9
Figure 7. A triangle test questionnaire using SIMS sensory software.....	10
Figure 8. Directional or attribute-specified difference tests.....	12
Figure 9. Unspecified or overall difference tests	14
Figure 10. An example of a triangle test case	19
Figure 11. Triangle test articles in the late 1940s until the late 1990s.	20
Figure 12. Sources of errors and their suggested solutions in triangle test	20
Figure 13. Articles on discrimination tests from year 2000s.....	21
Figure 14. Two possible options for triangle test.....	22
Figure 15. The degree of difference between products, δ	24
Figure 16. An example of the milk’s sample tray.	37
Figure 17. Results of discrimination tests with milk samples	42

Figure 18. Method comparison's results - Milk.....	43
Figure 19. Results of discrimination tests with juice samples.....	44
Figure 20. Method comparison's results - Juice.....	45
Figure 21. Results of discrimination tests with yogurt 1 samples.....	46
Figure 22. Method comparison's results - Yogurt 1	47
Figure 23. Results of discrimination tests with yogurt 2 samples.....	48
Figure 24. Method comparison's results - Yogurt 2	49
Figure 25. Results of discrimination tests with gurt samples.....	50
Figure 26. Method comparison's results – Gurt.....	51
Figure 27. Results of discrimination tests with cheese samples	52
Figure 28. Method comparison's results - Cheese	53
Figure 29. Summary of all the percent correct answers and computed p-values	56
Figure 30. Duo-trio vs. triangle and tetrad vs. triangle	57
Figure 31. Effects of series serving position to percent correct results	58

List of Tables

Table 1. List of available discrimination test methods.

Table 2. Steps in conducting a difference test.

Table 3. Advantages and drawbacks of the Guessing and Thurstonian models

Table 4. Selected discrimination testing methods for comparison

Table 5. Amount of sample servings, cups and time needed in each method

Table 6. List of products used as stimuli.

Table 7. Participation of assessors in all six sessions.

Table 8. Number of assessors who participated in each session

Table 9. Number of samples to be evaluated in each difference test series

Table 10. An example of given number codes to milk samples in each series

Table 11. An example of the design block showing the sample serving arrangements in each series.

Table 12. Steps in the evaluation part

Table 13. An example of a data sheet with topline results

Table 14. Minimum number of correct responses needed to establish significance at probability level of 5 % for duo-trio, triangle, and *tetrad* test

Table 15. Method comparison's open comments - Milk

Table 16. Method comparison's open comments - Juice

Table 17. Method comparison's open comments - Yogurt 1

Table 18. Method comparison's open comments - Yogurt 2

Table 19. Method comparison's open comments - Gurt

Table 20. Method comparison's open comments - Cheese

Table 21. Results of the three difference tests methods

Table 22. Summary of percent correct answers, p-values, and confidence levels

Appendices

Appendix 1 Justifications on method selection for comparison

Appendix 2 Test questionnaires

Appendix 3 Additional results

Appendix 4 Overall feedback

Appendix 5 Participant background

Appendix 6 Sample tray pictures

1 Introduction

In food industry, aside from the routine product quality control, products undergo different stages of development and modification to meet the constantly changing trends and consumer demands. Reasons for process modification and product reformulation can be due to cost-saving in production, recipe optimization, or compliance to sustainable development. There are also cases where alternative ingredients must be tested to secure the supply chain whenever problems arise from supplier change, shortage of raw materials due to environmental impact or interference in logistics. Due to economic reasons, the production can also be transferred from one manufacturing plant to another for processing efficiency. Recently there are increasing demands for testing alternative product packaging towards more environmental-friendly options to lessen the use of plastics materials. Implementing these modifications can create unwanted changes in the sensory characteristics of the product, which may result to negative feedback from the current product users. Even though the modification's aim is to meet a certain objective, it is important to make sure that the current sensory profile of the product will be maintained and the change will not be a risk. To verify whether the changes affect the product's sensory quality, discrimination testing can be conducted.

Discrimination tests are sensory methodologies used to determine whether differences between two confusable products are detectable by the assessors (Worch & Delcher, 2013, p. 396). Discrimination test means testing the ability to differentiate between two stimuli (Lawless & Heymann, 2010, p. 101).

Testing for the sensory difference between two products are routinely used in the industries. For determination of difference between two confusable products, the commissioner of this study, Valio Ltd. has been conducting triangle tests as the main method of discrimination testing both in product development and quality control. The triangle test is one of the most common discrimination testing method widely used until now. Results of triangle tests are used in the company to support different decision making in product and business development. Theoretically discrimination tests are simpler and faster to implement than other more detailed sensory evaluation methods such as consumer acceptance or

preference tests and trained panel's descriptive tests. In practice, organizing the triangle test is both challenging and time-consuming as it involves hours of sample cups coding, uniform sample portioning and random sample arrangement in balanced order combinations among the number of participants. Moreover, the number of samples needed for each product is almost twice more than the other sensory tests require. In addition, gathering enough participants needed for the test is always a challenge. In conducting discrimination tests, the participants are recruited among the internal panel who are willing to participate and evaluate the samples voluntarily. As an employee, participation to internal sensory tests in addition to the already busy working schedule is not always easy. To keep the employees motivated, participants are offered a take-away snack from the selection of the company's own products after the test. Although a small compensation is given after participation, there were occasions that the required number of participants is not met. Along with these challenges from the participant recruitment to test implementation, the test method itself is not flawless.

There are other types of discrimination testing methods available aside from the triangle test, some show advantages over the others while some showed limited application. Discrimination testing methods are still widely studied and reviewed until now. Recently, triangle test has received mixed reviews from other experts for being not stable and prone to many errors (J. M. Ennis & Jesionka, 2011; O'Mahony, 1995; O'Mahony & Rousseau, 2003). Currently, there is a wide range of literature and research published discussing which method is better than the other. There are also some modified versions of the different discrimination methods to answer specific objectives or to increase the test method's power and sensitivity (Jeong et al., 2016; Kim et al., 2014b; Rousseau & O'Mahony, 2000; van Hout, 2014; Xia et al., 2015). With recent studies on other discrimination testing methods found to be more powerful than the triangle test, alternative methods for the triangle test must be reviewed. The stimuli used in previous studies were mostly simple solutions and only several studies have tested different methods using actual food samples. As most of the studies' recommendation, testing these methods with the company's own range of products will be beneficial to better understand the applicability of each method on different test cases.

2 Aim of the study

The aim of this thesis was to investigate the reliability and sensitivity of the currently used method by comparing other similar discrimination methods available and recently found to have advantages over the triangle test. The two main research problems in this experiment were:

1. Is triangle test still the best sensory discrimination method for difference testing in the company?
2. If not, what are other options available?

Investigating meant diving deep into the vast waters of scientific papers on the recent reviews and developments in discrimination testing. After being familiarized with other methods' potential, a direct application and comparison to triangle test was implemented using the company's own product range. In addition to the main research problems, this study aimed to answer these specific research questions as well:

- a. What are the food industries' widely used triangle test's strengths and weaknesses?
- b. What are other similar discrimination testing methods available, and their advantages or disadvantages compared to the triangle test?
- c. Can the current method be replaced or partly substituted by other similar methods?

The first two research questions (a. and b.) will be answered in the literature review, then the remaining research question (c.) and research problems (1. and 2.) will be answered after the experimental part.

A theoretical background and literature review will be provided first in the next sections, before moving on to the experimental part.

3 Sensory science

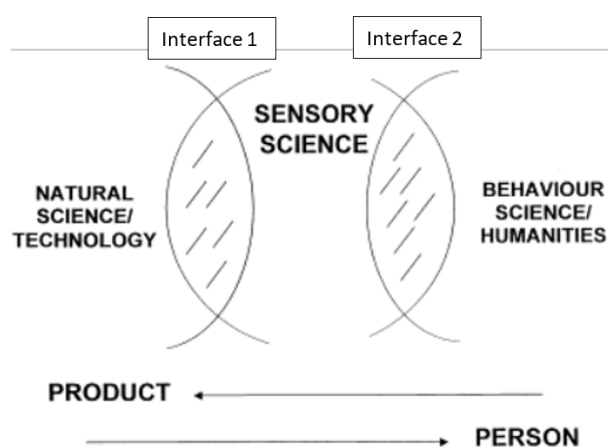
Sensory science is defined by Martens (1999, p. 233) as “a multidisciplinary field comprising measurement, interpretation and understanding of human responses to product properties as perceived by the senses such as sight, smell, taste, touch and hearing.” The same author (Martens, 1999, p. 234) also elaborated the field of sensory science in a philosophical way:

Sensory science is

- i. *relational, i.e. links product and person;*
- ii. *interdisciplinary, i.e. links professions from chemistry to psychology;*
- iii. *timeless, i.e. links past and present inquiries into sense perception;*
- iv. *existential, i.e. links sense perception through one human life;*
- v. *integrating, i.e. links the various sense modalities;*
- vi. *“real world” science, i.e. links theories to practical problem-solving.*

An illustration (Martens, 1999, p. 235) below in Figure 1. shows how sensory science studies both product–person relations (Interface 1) by interpreting chemical-sensory properties and person–person relations (Interface 2) by combining chemistry, psychology, and marketing to interpret descriptive-affective responses.

Figure 1. Sensory science as a “link” to other fields of science



3.1 Sensory evaluation

Sensory evaluation is defined by Lawless & Heymann (2010) as “a scientific method used to evoke, measure, analyze and interpret reactions to those characteristics of foods and materials as they are perceived by the senses of sight, smell, taste, touch and hearing”. In a sensory evaluation, a stimulus (product) is presented to a subject (human) and the subject’s reaction to the stimulus (response or sensory perception) is measured. Human perception to the product tested is converted numerically for statistical analysis. In other words, sensory evaluation is performed by the human senses either as “measuring instruments” during an analytical laboratory sensory evaluation or as “predictors” during consumer in-house or in-hall sensory evaluation. These types of evaluation are defined by O’Mahony (1995) as Sensory Evaluation I (SE I) using trained panelists as assessors and Sensory Evaluation II (SE II) using untrained consumers as assessors. More on the definition of both types of sensory evaluation can be found from the studies of O’Mahony and Hout (1995; 2014).

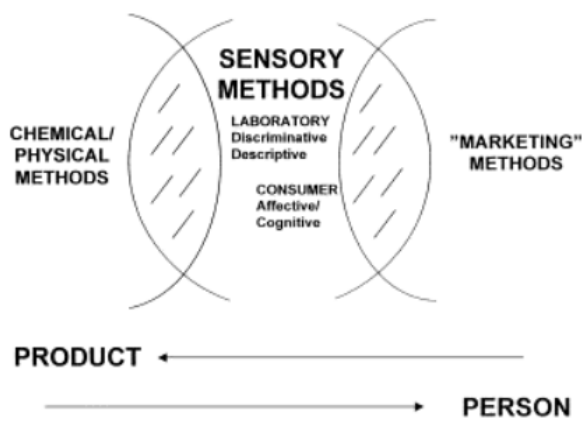
Sensory testing is used to gather information on product properties to improve, modify and maintain the product quality. Sensory evaluation aids the research and development process in industrial manufacturing companies to fully understand their products and assess consumer response. (Amerine et al., 1965)

As we all know humans are complicated beings prone to inconsistencies and no matter how much “calibration” and training one undergoes to provide the most accurate sensory responses to a product being tested, human’s judgement is often influenced by many circumstances. These creates variations in the results. Although accurate measurements may not be fully achieved by sensory evaluation by humans, there are currently no other more accurate measurements available to measure human’s sensory responses to consumer goods. Recent developments have been reported and achieved in sensory evaluation by actual instruments with the help of modern technologies and artificial intelligence such as electronic (e-) nose, e-tongue and e-eye as suggested by recent studies (Crofton et al., 2019; Fuentes et al., 2021; Gonzalez Viejo et al., 2019; Motoki et al., 2021; Ross, 2021), but none up to this day can combine all the e-senses into one holistic measurement of sensory perception by human.

3.2 Sensory evaluation methods

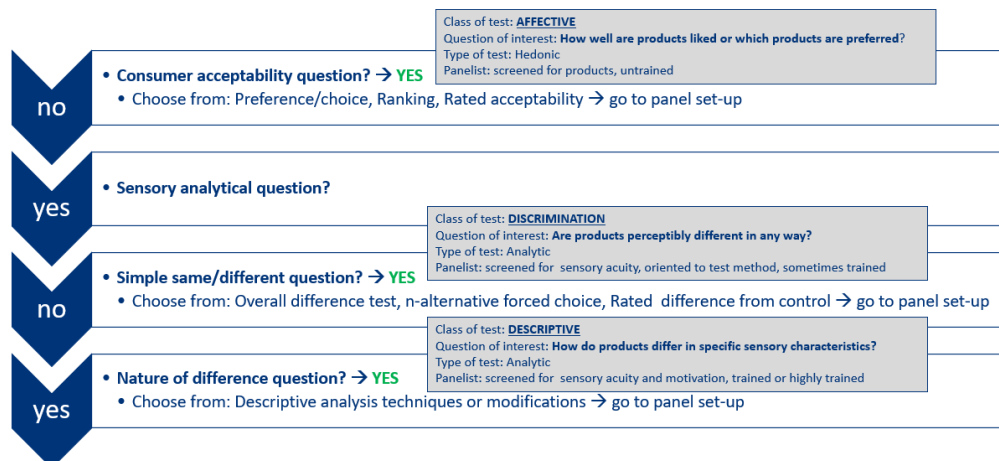
There are different sensory evaluation methods available depending on the objectives of the study and what research questions must be answered. The main methods of sensory evaluation are commonly divided into two categories: laboratory tests and consumer tests. A concise illustration of sensory evaluation methods was described by Martens (1999, p. 240) in Figure 2. below.

Figure 2. Main methods of sensory evaluation



A detailed illustration on how the different sensory evaluation methods answer different research questions was also described in the book of Lawless & Heymann (2010, p. 16) as modified in Figure 3.

Figure 3. Method selection based on research objective

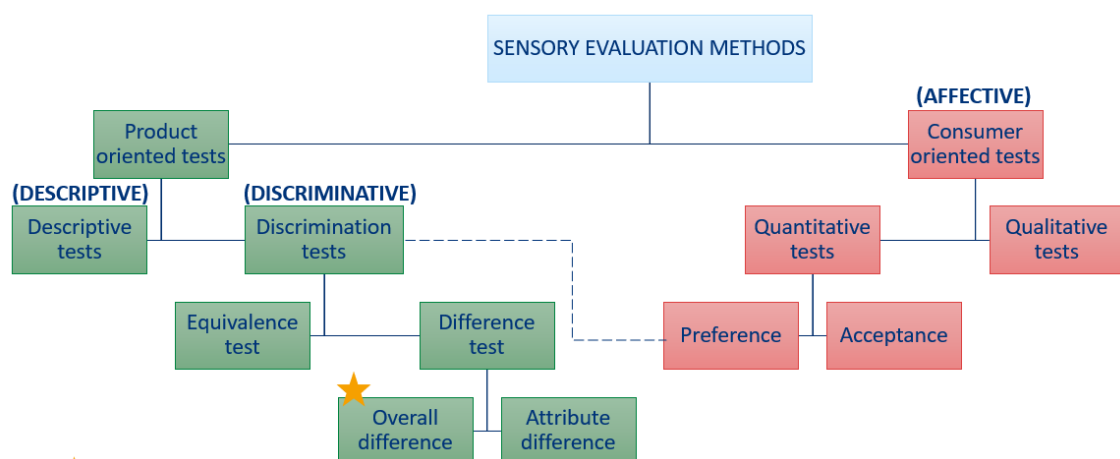


The sensory evaluation team in a food manufacturing company mainly supports research and development but interacts with marketing, quality control, packaging, design, advertising, legal, and regulatory teams as well (Lawless & Heymann, 2010, p. 16). It is the task of the sensory scientist to choose and design the most appropriate sensory method to answer the research questions needed for product development, business, or marketing decisions.

There are three main classes of sensory testing as briefly described in Figure 3: affective, descriptive, and discriminative. All three serve a different purpose and provide companies with different set of information to answer different set of research questions.

Illustrated below in Figure 4 are sensory evaluation methods grouped into two focus areas: product-oriented tests and consumer-oriented tests (Adjei, 2017, p. 86). Product-oriented tests, which Martens previously referred as laboratory tests, include the descriptive tests and the discriminative tests. Consumer-oriented tests include affective tests, which can be done quantitatively using preference test and acceptance test or qualitatively using interviews and panel conversations. The line between the two focus areas is very distinct, but recent studies showed the incorporation of affective preference to discrimination tests allowing an invincible line connecting the two focus areas be drawn (Kim et al., 2014, 2015). The yellow star indicates the method of focus in this study.

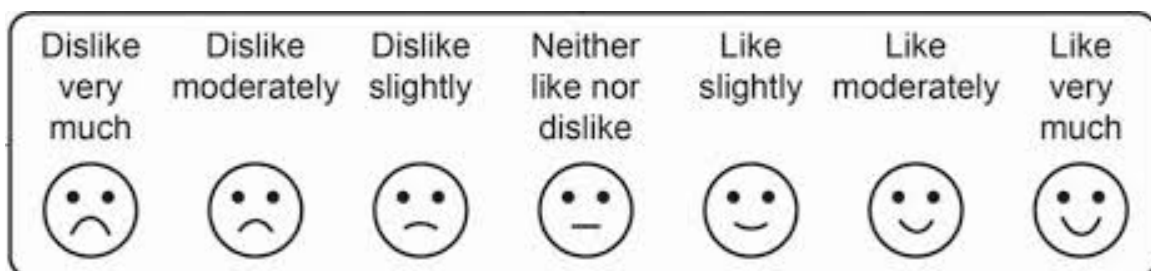
Figure 4. Product-oriented tests vs. consumer-oriented tests



Affective testing

For research questions such as, “How is the product liked or not liked?” or “After tasting the new product, will the consumers accept the product and intend on purchasing if made available in the market?”, an affective type of testing must be conducted. Affective or hedonic testing measures the consumers’ likes and dislikes to determine the product’s market potential. This method requires a group of representative consumers ranging from about 50-100 persons, who must represent the actual or probable end-users of the products to be tested. To take part in this type of sensory evaluation for product’s acceptance and preference, the participants must be screened for motivation and product use. The evaluation must take place in exactly or as near as the actual product consuming situations in the real world. The participants do not require prior training (untrained panelists) and the location is not centralized. This type of test is often referred to as home test, which involves ordinary consumers without prior sensory evaluation training, performing the test in the convenience of their own homes. Affective testing can also be performed in a controlled manner in a centralized location, like the sensory laboratory or even in a bigger testing hall to accommodate larger number of participants. This type of testing is also known as hall test, which involves ordinary consumers without prior sensory evaluation training, but not in the convenience of their own homes. This allows the sensory scientist to limit the variations of the affecting variables like location, product presentation and manner of evaluation. Acceptance is measured by various types of hedonic scales. Shown below in Figure 5. (Abdou et al., 2018, p. 290) is an example of a seven-point facial hedonic scale for measuring product likeability.

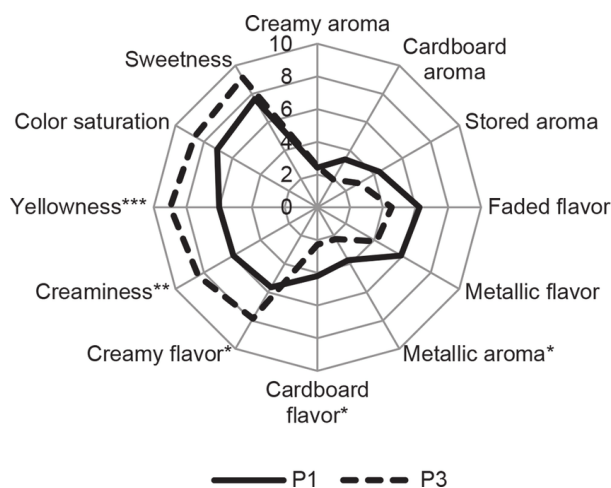
Figure 5. Seven-point facial hedonic scale



Descriptive testing

Descriptive testing on the other hand is used when a product's sensory characteristics or attributes must be specified and described in detail. This type of testing answers the questions like, "What is the nature of the products' difference? or How do products differ in certain sensory characteristics and their intensities? The evaluation is commonly performed by a trained group of product experts as panelists in a controlled testing facility like the sensory laboratory. A training session prior to the descriptive testing is organized to define which sensory attributes will be evaluated and the range of scale to be used for measuring each attribute's intensities. The range of scale depends on how different the samples are based on their sensory qualities. During the training session, the panelists are presented with reference samples to be evaluated. After tasting, each panel member lists all the sensory attributes, where the samples differ in intensities. A discussion on the attributes of concern will follow, then the range of scale will be decided based on consensus. After training, each will perform the descriptive analysis of all the samples individually. Evaluation is done analytically focusing mainly on sensory characteristics, while trying to exclude personal preference. An example on how to report the results of a descriptive analysis by a cobweb diagram, is shown in Figure 6. using a ten-point scale (Maciel et al., 2016, p. 8533). The diagram easily visualizes the results showing how the two samples P1 and P3 differ from each other and what sensory attributes are missing or present in each sample. The asterisks beside the attributes show the level of significant differences between the samples.

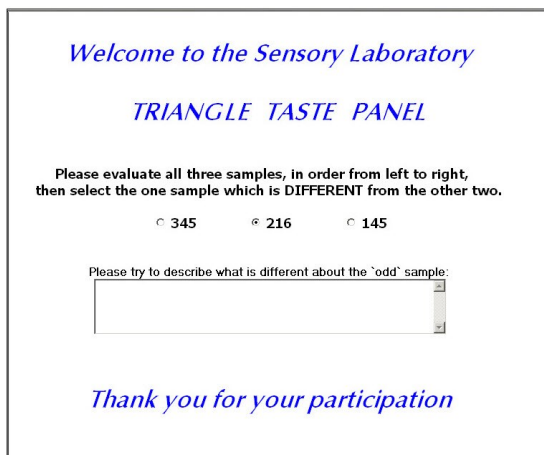
Figure 6. A sensory profile of two milk samples



Discriminative testing

The third type of sensory evaluation method is the discriminative or better known as discrimination testing. This method is used to measure small differences or similarities between two products. Difference and similarity tests can be used to identify whether an overall or attribute-specified difference or similarity exists. This type of sensory testing answers research questions like “*Are the two products perceived as different?*” or “*Does the product’s processing modification result in a perceptible sensory difference from the current version?*”. The panelists can be trained specialists or untrained consumers. Choosing which panel to use depends on the objective of the research. A case study comparing the trained and untrained panels’ performance in difference testing were presented in the dissertation on measuring meaningful differences by van Hout (2014, pp. 26–32). Evaluation in discrimination tests is also analytic like the descriptive test, meaning no product preference or acceptance questions are involved. Although there are cases combining both affective and discriminative tests as a modification to meet specific needs, the main aim of any discrimination tests is to answer the basic question: Are the products perceived different or not? There are numbers of discrimination testing methods available, each has its own strengths and weaknesses. One example is the triangle test, which is the main subject of this study. An example of a triangle test questionnaire using an online cloud service, SIMS sensory software (SIMS, n.d.) is shown in Figure 7. More details on discrimination tests will be discussed in the next part.

Figure 7. A triangle test questionnaire using SIMS sensory software



Welcome to the Sensory Laboratory

TRIANGLE TASTE PANEL

Please evaluate all three samples, in order from left to right, then select the one sample which is DIFFERENT from the other two.

345 216 145

Please try to describe what is different about the “odd” sample:

Thank you for your participation

4 Discrimination testing methods

The literature review focused on discrimination tests, more particularly in difference testing. The task was to learn and understand more about the different methods available, which can be possible alternatives or substitutes to triangle test. In this part, the research questions regarding the triangle test's strengths and weaknesses compared to other methods will be discussed along with other available methods for consideration.

As mentioned earlier, products can undergo reformulation, processing modification, switch to an alternative packaging material, extension of shelf-life or adjustments in storing or serving conditions. Before any of these changes to be officially implemented in the company, the product development and business management teams must test and confirm how these changes will affect the sensory characteristics of the product, and if there are perceivable differences, how much change will be agreed to be acceptable. To determine whether there are perceivable differences between the new or modified product (B) and the current version of the product in the market (A), sensory discrimination tests are used to avoid compromising the products already liked by the consumers.

Discrimination testing is only applicable if the two products to be tested are confusable, meaning the difference in their sensory quality is difficult to distinguish clearly. This sensory testing method will measure the effect of product modification by comparing the prototype (modified product) to the control (reference product). The control can be the current version of the product available in the market, a standard sample, or a target product with sensory characteristics to be achieved.

There are different kinds of discrimination tests available depending on the test objectives of the study. It is the responsibility of the sensory scientist to recommend and choose the most suitable testing method for the products. Difference testing is more commonly used than equivalence or similarity testing, because the number of required participants is usually higher when testing for similarities. Although both types of tests seem interchangeable, the data analysis and interpretation of results are different. The technicalities on how difference and similarity testing differ from one another are explained in the sensory analysis standards

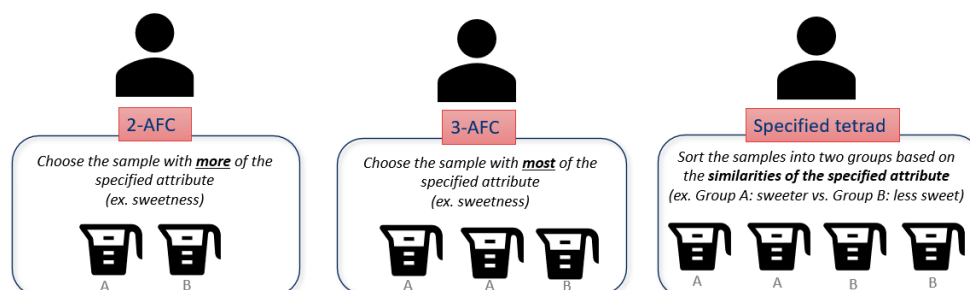
(ISO 4120:2021, 2021; ISO 6658:2017, 2017). Since difference testing is more common and is what the company has been conducting, this study focused on this type of discrimination test. Some of the available methods include the triangle test, duo-trio test, paired comparison test, n -alternative forced choice test (n -AFC), tetrad test (Frijters, 1984, pp. 117–140), polygonal, and polyhedral tests (Basker, 1980, pp. 1–10).

These test methods can be further divided into two groups: overall difference testing and attribute-specified difference testing. More about the differences between specified and unspecified testing methods will be discussed below.

1. Directional difference testing (specified)

If product's attribute subjected to change is known or specified, there are attribute-specified discrimination testing methods available such as the n -alternative forced choice (n -AFC) test like 2- or 3-AFC. Respondents are presented with two to three samples and are instructed to identify which sample has more or most of the pre-specified attribute (J. M. Ennis & Jesionka, 2011). Tetrad test can also be presented as a specified version, in which test instruction is different from the other two; the assessor is presented with four samples that must be sorted into two groups based on similarities according to the specified attribute (Xia et al., 2015). Illustrations on how samples and testing instructions are presented in these methods are shown in Figure 8.

Figure 8. Directional or attribute-specified difference tests



This type of discrimination testing is applicable on situations like sugar or salt reduction when a company aims to lessen the sugar or salt content of the product, without changing

the sensory taste profile drastically. In this case, the attribute subjected to change and is of main concern is sweetness or saltiness. Using specified difference tests such as 2-AFC, asking the assessor specifically which one between the samples is sweeter or saltier, the discrimination task is straightforward and is easier to perform. When the attribute of concern is defined, it becomes easier for the assessors to focus on their task in discriminating the samples based on only one sensory characteristic. This is the reason why specified discrimination test is more effective and powerful than the unspecified difference tests.

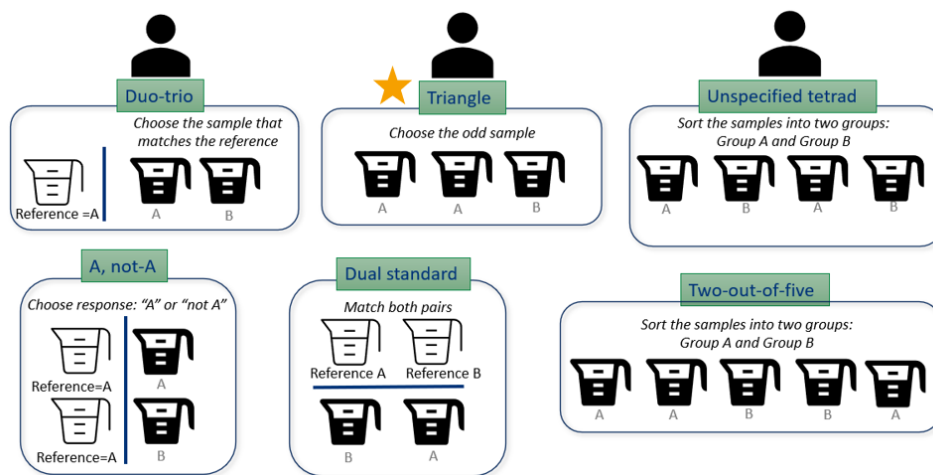
Although specified methods are much more powerful than unspecified methods like duo-trio, triangle, unspecified tetrad, and two-out-of-five (Bi, 2015; J. M. Ennis & Jesionka, 2011; McClure & Lawless, 2010; O'Mahony & Rousseau, 2003), very seldom that a company can predict a certain attribute change resulting from modifications in processing, packaging, or ingredients. Thus, testing for overall differences between samples undergoing product modification is more applicable to the company's purposes. Although a specified difference test would be a better choice for testing in some cases, overall difference testing is applicable in most cases being the safest and general method of choice. This was the reason why specified methods, although being more accurate than overall methods, were not included in the comparison.

2. Overall difference testing (unspecified)

When the sensory attribute subjected to change is unknown or if the modification will affect several sensory attributes, unspecified or overall difference testing must be conducted. There are several overall difference testing methods available such as A-not-A, dual standard, duo-trio, triangle, unspecified tetrad, and two-out-of-five. All these methods involve presenting the respondents with two to five samples. The instructions vary depending on the method used. In A-not-A, a reference sample A is presented, then the assessor is asked to state whether the following test sample is A or not-A. On the other hand, dual standard presents both samples as references A and B, then the assessor must match the two test samples to the references. In duo-trio test, the respondents are presented first with the reference sample, followed by two samples. The tasks are to taste

the reference first, then identify which of the two samples is the same as the reference. There are also two versions of duo-trio test: one uses the same reference sample throughout the evaluations (constant reference) and the other uses both samples as the reference sample in a random order (balanced reference). Using the constant reference duo-trio test provides simplicity in sample preparation and is appropriate if the assessors are familiar with the reference. In the company, the reference sample used in most of cases was the current product available in the market, which was already familiar to the internal panel. By using a familiar product as reference, the matching task during the evaluation is easier. In cases, where both samples to be tested are unfamiliar to the assessors, for example new novelties for future market launch, the balanced reference duo-trio test is recommended. In the triangle test, among the three samples, the task is to identify the odd one. In unspecified tetrad and two-out-of-five tests the task is to sort the samples into two groups depending on similarities. Figure 9 shows the sample and instruction presentations in each test. The yellow star mark indicates the current method used in the company.

Figure 9. Unspecified or overall difference tests



These methods are suitable for testing overall difference, meaning there is no specified sensory characteristic or attribute known or defined prior to the test implementation. For example, a company wanted to change the supplier of cheese starter for cost-reduction. As a trial, the product development team made a test cheese sample using the same starter but from the new supplier. To verify if the supplier change does not affect the sensory quality of the cheese, the test sample will be compared to the current version available in the market.

Since cheese processing is complicated and predicting what sensory attribute is going to change using the new starter, overall difference test would be the suitable testing method to be used. By doing so, the tester (assessor) will be asked to identify which among the three samples is the odd one (triangle test) or which one between the two samples matches the reference sample (duo-trio).

The summary of different discrimination test methods modified from the book by Lawless and Heymann (2010, p. 81) is shown in Table 1. The newest method called the tetrad test was not yet included in the book but is listed here as it grows popularity in recent studies showing high potential in replacing other discrimination testing methods, like the triangle test. Tetrad was found to be more powerful than the triangle test due to fewer required assessors to achieve reliable results (J. M. Ennis & Jesionka, 2011, p. 381).

The table shows how wide the selection of different discrimination testing methods currently is, and by just looking at the probability column values, one can easily see that the method with the least chance of guessing the correct answer (probability, p) like two-out-of-five ($p=1/10$) or 4/8 "Harris-Kalmus" ($p=1/70$) will be the most ideal choice. Although the probability of guessing the correct answer by chance in these methods is low, the number of samples included in the evaluation set-up increases, thus causing an increase in tasting fatigue. With strong and complexly flavored samples, tasting more samples will be difficult to the assessors and the carryover of lingering tastes from one sample to the next will adversely affect the discrimination task. These methods will be more appropriate to sensory evaluation of color or appearance, but not when the evaluation includes tasting and texture evaluation through mouthfeel. Since taste is one of the most important sensory characteristics valued by the consumers in addition to the texture and appearance, the method must also be applicable to all the sensory quality evaluation of the product. Aside from the difficulty in the discrimination task caused to the assessors, more samples will require a greater number of products and time needed for test implementation. Therefore, the advantage of providing low probability of guessing the correct answer by chance is overpowered by the disadvantages in evaluation fatigue, time consumed in test preparation and sample amount needed for the test. With these reasons being stated, the methods with low probability will not be a better alternative for triangle test.

Table 1. List of available discrimination test methods.

CLASS OF TEST	NAME OF TEST	PRE-TEST SAMPLES	TEST SAMPLES	TASK/INSTRUCTIONS	PROBABILITY
Oddity	Triangle	None	A, A', B or (A, B, B')	Choose the most different sample	1/3
Matching	Constant reference duo-trio	Ref A	A, B	Match sample to reference	1/2
	Balanced reference duo-trio	Ref A, Ref B	A, B	Match sample to reference	1/2
	ABX	Ref A, Ref B	A (or B)	Match sample to reference	1/2
	Dual standard	Ref A, Ref B	A, B	Match both pairs	1/2
Forced choice	Paired comparison	None	A, B	Choose sample with most of the specified attribute	1/2
	3-AFC	None	A, A', B	(Same)	1/3
	n-AFC	None	A ₁ -A _{n-1} , B	(Same)	1/n
	Dual pair	None	A, B and A, A'	Choose A, B (different pair)	1/3
Sorting	Tetrad	None	A, A', B, B'	Sort into two groups	1/3
	Two-out-of-five	None	A, A', B, B', B''	Sort into two groups	1/10
	4/8 "Harris-Kalmus"	None	A ₁ -A ₄ , B ₁ -B ₄	Sort into two groups	1/70
Yes/No	Same-different	None	Pairs: A, A' or A, B	Choose response "Same" or "Different"	N/A
Response choice	A, not-A	Ref A	A or B	Choose response "A" or "not-A"	N/A

To show the whole process, the steps in conducting a discrimination test are enumerated in Table 2 (Lawless & Heymann, 2010, p. 81).

Table 2. Steps in conducting a difference test.

1. Obtain samples and confirm test purpose, details, panel, training, and client.
2. Decide testing conditions such as sample size (number of participants), volume of samples, serving temperature and confirm with the client.
3. Write instructions to the panelist and construct ballot (test questionnaires).
4. Recruit potential panelists.
5. Screen panelist for acuity (if necessary).
6. Train to do specific difference test (if necessary).
7. Set up counterbalanced sample orders.
8. Assign random three-digit codes and label sample cups or plates.
9. Conduct test.
10. Analyze results.
11. Communicate results to clients or end user.

The digitalization era has also changed the sensory evaluation from paper questionnaires to internet-based questionnaires making these steps faster to accomplish. Conducting a sensory test automated from the recruiting part to analyzing the results in one platform was made possible using advanced sensory software and web applications. The only part, which is excluded from the automation, is the sample preparation and the sensory evaluation itself. These are the areas where considerations in choosing the appropriate method can be found: the lower number of samples and participants needed, the lesser time required to conduct the test and the easier the discrimination task will be for the assessors, the better and more practical the method is.

The currently used triangle test has a long history of use behind. It is a popular choice for discrimination testing and has provided useful results to support business decisions not only in Valio, but in other companies as well. But what is wrong with triangle test that the need to investigate other methods was necessary? It has always been criticized in several publications, but it was not looked upon closer yet in the company.

4.1 Triangle test and its pitfalls

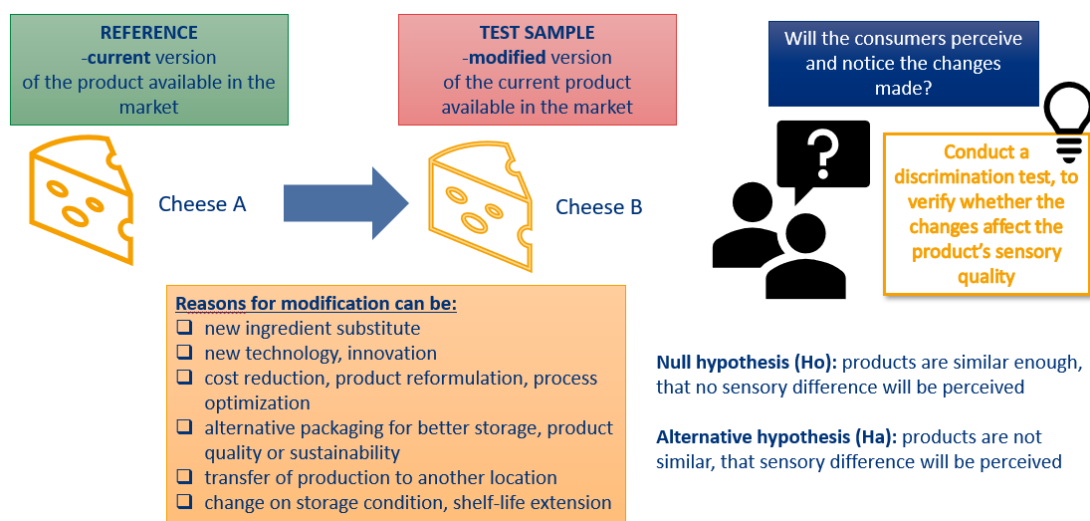
Now that many other methods aside from the triangle test were listed, the next step of this study was to choose which among these methods were to compare. Due to the limitations of this study, it was decided to focus mainly on comparing similar methods like the one currently and actively used in the company: the triangle test.

Triangle test is one of the most commonly used discrimination testing method due to its theoretical simplicity both in test implementation and data analysis. Results are also easy to understand and communicate to the test requesters. Triangle test is a type of discrimination testing method used to determine whether an overall difference between two samples is perceivable. Overall means that the difference can be in appearance, texture, taste, smell, or all combined. It is also used to train and screen the panel involved in product quality assurance such as testing for off-flavors identification or taste sensitivity thresholds. In triangle test as shown briefly in the previous discussion, a prototype (B) is compared to a reference (A). The reference can be the current version of a product or a standard sample with the target sensory qualities. The prototype can be the modified or reformulated sample. The laboratory test is organized by presenting the samples in a unified manner: same sample serving container, size, amount, appearance, and temperature. Each test participant will be presented with three samples coded with three-digit numbers. The samples must be randomly arranged among the participants to make sure all the sample combinations AAB, ABA, BAA, BBA, BAB, and ABB are served. The tasks will be to taste all the samples according to the order presented in the sample tray or questionnaire, then to choose the odd one. The odd sample can be either the reference or the prototype. If the difference between the samples is not perceived, the participant must guess. If the total number of correct answers from all the participants is less than the minimum number of correct answers needed to establish statistical significance, it will therefore be concluded that there is no significant evidence, that perceivable difference between the two samples exists. (ISO 4120:2021, 2021)

Theoretically, test implementation, sample evaluation, data analysis and interpretation of the results are faster and easier compared to the other sensory tests such as the acceptance

or descriptive tests. In practice, the number of samples, amount of time and other resources involved in conducting a triangle test are not that simple. Moreover, the reliability of the triangle test is often criticized as shown in several articles published (D. M. Ennis, 1993; J. M. Ennis & Jesionka, 2011; O'Mahony, 1995b; O'Mahony & Rousseau, 2003). Below in Figure 10. is an example of some actual situations in a food industry, where product modification is needed but the change in sensory quality must be verified. This can be an actual case, where triangle test is applicable.

Figure 10. An example of a triangle test case



Triangle test is the most common method in the food industry and is also the current method used in the company. The results provided important support in decision making and the method has been used for over thirty years in the company.

Although triangle test came in the late 1940s (Roessler et al., 2006), it has been since then criticized, reviewed, and modified in several books and articles published up to date. According to a presentation of Tom Carr in the recent Pangborn Symposium (2021), there are currently 259 scientific papers published in the Journal of Sensory Science about the triangle test alone. Figure 11. below shows just a pair of published articles about the same topic, but if looked closer at their dates, the time difference was almost half a decade.

Figure 11. Triangle test articles in the late 1940s until the late 1990s.



A comprehensive review done by O'Mahony (1995) discussed the effects of position bias, response bias, cognitive strategy changes and the sequence of tasting in relation to the triangle test (Figure 12.). He also elaborated on the theoretical approaches like 'Thurstonian modeling and Sequential Sensitivity Analysis'. His study concluded that the triangle test is prone to many pitfalls (O'Mahony, 1995, p. 236) as shown in the figure below.

Figure 12. Sources of errors and their suggested solutions in triangle test

Position bias	<ul style="list-style-type: none"> randomizing or counterbalancing the order of presentation of the stimuli in the triad
Order of tasting	<ul style="list-style-type: none"> triad with the stronger stimulus as the odd increase the sensitivity for better discrimination
Response bias	<ul style="list-style-type: none"> instructions must indicate the number of stimuli on each side of the criterion.
Changes in cognitive strategy	<ul style="list-style-type: none"> prevent the switch from overall comparison of distances (triangle) to directional skimming (3-AFC) strategy by constantly changing the stimuli in the test

In the article review by O'Mahony (1995), the sources of variations in the results from a triangle test in relation with the position bias, can be controlled by counterbalancing the sample randomization among the participants to make sure all the possible sample arrangements AAB, ABB, ABA, BAB, BBA, and BAA are presented equally. To avoid response

bias or choosing a specific response more often than the other, the test instructions must be written carefully and the task and number of stimuli must be stated clearly in a straightforward manner. The concerns in the order of tasting, with series containing the stronger stimulus as the odd sample resulting to better discrimination than those who received a series with the weak stimulus as the odd sample is presented. Also, in performing the triangle test, the assessor can switch between cognitive strategies either discriminating the samples based on overall perception or by skimming based on a specific attribute as the cognitive strategy used in directional difference testing like the paired comparison, 2-AFC.

During these recent years, more articles are being published not only about triangle test, but also in discrimination tests in general. This means that in addition to the already wide range of articles to be reviewed, the list still goes on. Figure 13 shows some of the recent studies on modification, combination, and improvisation of several discrimination testing methods (In-Ah Kim et al., 2015; Ishii et al., 2014; Jeong et al., 2016; Kim et al., 2014a; Kuesten, 2001; O'Mahony & Rousseau, 2003; van Hout, 2014; Xia et al., 2015).

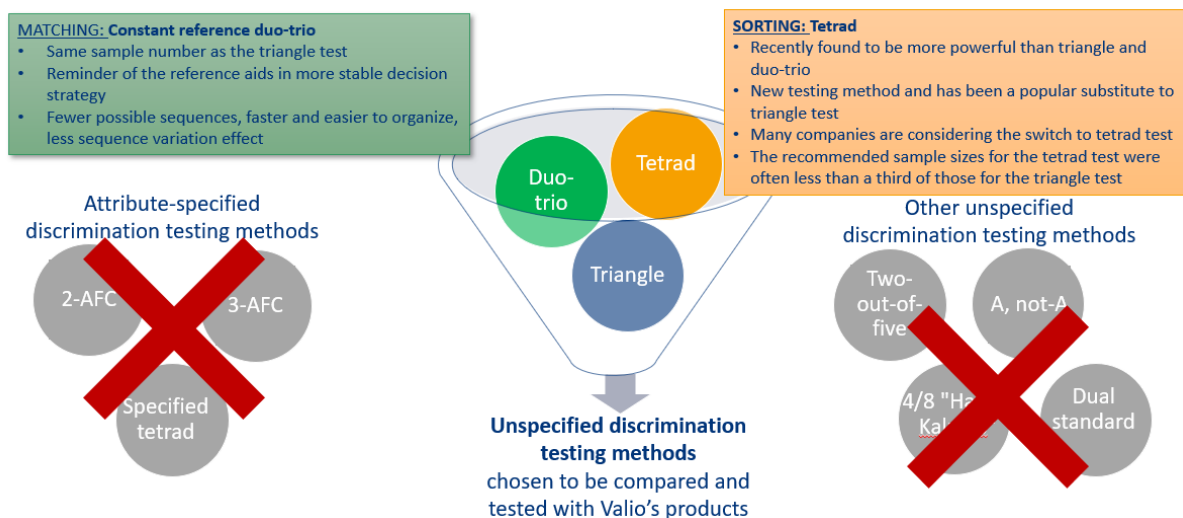
Figure 13. Articles on discrimination tests from year 2000s



For the purpose of this study and the limitations agreed upon, the main goal was to focus on testing and comparing two other overall difference testing methods with the currently used triangle test. These two methods agreed were the duo-trio test and the unspecified tetrad test as shown in Figure 14. The justifications on the selection of methods to be compared are included in the Appendices section. Constant reference duo-trio test might not be as popular

as the triangle test but is more straightforward and both easier to organize and perform than the triangle test. The higher chances of guessing the correct answer (1/2 probability) in duo-trio test may lessen the effectivity of the method but was intended to test and confirm in this study. Although the tetrad test is a newcomer, it had gained attention and earned interests for discussion among sensory experts as a better choice than triangle test. Other companies are now in transition and considering the switch from triangle to tetrad (Bissmeyer, 2019; J. M. Ennis, 2012) There are several comparisons made between triangle and tetrad tests as written in several recent publications (Adawiyah et al., 2020; Chaves et al., 2020; Garcia et al., 2012; Ishii et al., 2014; Theses & Carlisle, 2014; Tran et al., 2014). The addition of the fourth sample in tetrad test may increase sensory fatigue, that may affect the discrimination performance of the panel, which was later tested to confirm in this study as well. This experiment was designed to test and compare all the three methods in one testing session using some of the company's wide range of products.

Figure 14. Two possible options for triangle test



This study aimed to provide a reference for discussion whether the company will continue to use the current method or replace it with a more accurate and cost-time-efficient method.

4.2 Methods of analysis

Analyzing discrimination tests results can be done using two different models: the guessing model and the Thurstonian model. The guessing model is more commonly used because of its simplicity requiring only simple calculation and the results are easy to understand and interpret. The Thurstonian model is known for producing more precise results, but it is more difficult to understand and interpret. Although the two models use different approaches, Worch and Delcher (2013) were able to present a practical guideline for users in applying both models for analyzing the results.

1. Guessing model

The individual scores of the respondents in any discrimination tests are recorded as binary results: a correct answer scores 1 and a wrong answer scores 0. Individual scores are not used, but the total number of correct answers instead. The proportion of the total number of correct answers, P_C computed in relation to the total number of responses received is shown in Equation 1.

Equation 1

$$P_C = \frac{\text{Total number correct answers}}{\text{Total number of trials}}$$

A first way to analyze the data from a discrimination test is by estimating the proportion of discriminators, P_D using the guessing model (Meilgaard et al., 2006, pp. 63–65). This model assumes that the panel of assessors can be divided into two subgroups of assessors: the discriminators and the guessers.

The guessing model requires only simple calculations and is easy to understand and interpret. The proportion of discriminators, P_D can be computed using the proportion correct, P_C and the proportion of guessers, P_G as shown in Eq. (2):

Equation 2

$$P_D = \frac{P_C - P_G}{1 - P_G}$$

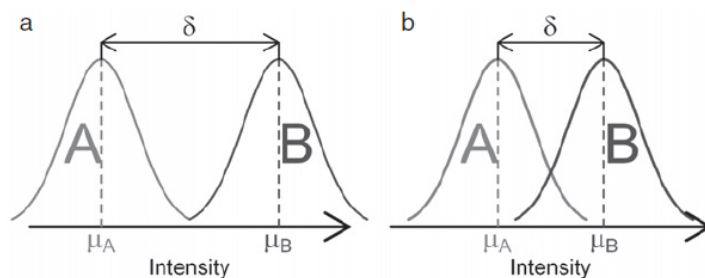
P_G depends on the guessing probability of the method in use. The proportion of discriminators and the proportion of correct answers are linearly linked (Ennis, 1993).

2. Thurstonian model

A second way to analyze P_C uses the signal detection theory first introduced by Thurstone (1927). The Thurstonian approach converts the total correct answers into an estimate of the sensory difference called d-prime, d' . Although the computation involved in this model is not straightforward, Worch & Delcher (2013) succeeded in explaining the theory behind this approach: “When two products are compared, the distance between the two products is measured by the parameter δ , which is the distance between the means of the corresponding perceptual distributions, weighed by the inverse of the standard deviation of the perceptual distributions. Thus, δ is the signal-to-noise ratio”.

Figure 15. shows that two products are easily distinguishable if the corresponding distributions are so distant that they are not overlapping (a). Similarly, two products are confusable if the corresponding distributions are overlapping (b). The more the distributions overlap, the more the products are confusable.

Figure 15. The degree of difference between products, δ



Worch & Delcher (2013) provided an excellent summary of the comparison between the two models as shown in Table 3.

Table 3. Advantages and drawbacks of the Guessing and Thurstonian models

	Guessing model	Thurstonian model
Calculations	Easy models based on the assumption that the subjects are either guessers or discriminators	Advanced models based on assumption that the product perception follows a certain distribution (often defined as the normal distribution)
	Subject related	Product related
	No need of a particular software	Necessity of a specialized software providing d'
Parameter estimated	P_G	d' (statistical estimate of δ)
	It corresponds to the true proportions of discriminators, i.e., the true proportions of subjects who perceive the difference between products	It corresponds to the distance between the mean distributions representing the perception of the products, and weighted according to the perceptual noise (signal-to-noise ratio)
	It is calculated using a linear relationship involving P_C and P_G	Psychometric functions relating P_C and δ for each protocol
Fixed protocol	Powerful if well defined	Powerful
Across protocols	Unstable (method-specific)	Stable (not-method-specific)
Interpretation	Easy to interpret and communicate	More difficult to interpret and communicate
	Existence of well-defined tables to help the users	Currently, no well-established table exist to support the users

The Thurstonian approach is recognized to be more stable across protocols than the guessing model, however it has been criticized due to the more complicated calculations

required (Lawless & Heymann, 2010). Moreover, the interpretation is less straightforward than for the guessing model.

Further studies have explained how the complexity of calculating d' is no longer an issue. A recent article by Christensen (2020) is published describing the statistical methodology for sensory discrimination testing and analysis, and how the analyses can be performed in R programming language using package `sensR` developed by Brockhoff and Christensen (2010).

In the article by Ennis and Jesionka (2011) "The power of sensory discrimination methods revisited" (PSDM), the need for power considerations in the interpretation of testing results is clarified and series of sample size tables are provided for easy interpretation. Sample size is also known as the number of participants.

Although these two models of data analysis and interpretation were interesting to investigate further and incorporate with this study, the limitation in time and resources did not allow some practical application with the company's products. The reason why these models were included in the literature review was to bring out the possibility of considering these options in data analysis. More about this topic was discussed in several papers listed in the reference section (Brockhoff & Christensen, 2010; J. M. Ennis et al., 1998; Linander et al., 2019; Thurstone, 1927; Worch & Delcher, 2013).

5 Materials and methods

This experiment tested and compared three different overall or unspecified discrimination testing methods in one session. One session means one product test using all three methods in series. The three test methods selected were the triangle test, the constant reference duo-trio test and the tetrad test. The justifications for choosing the selected methods for comparison were illustrated in detail in the Appendix section. All the reasons were based on the dissertation of Hout on measuring meaningful differences (2014) and the review made by Ennis and Jesionka (2011) on the power of sensory discrimination methods. Only few of the main product categories were selected to represent the wide range of product varieties due to the limitations imposed by the pandemic. In sensory discrimination testing, two

products or two versions of the same product are compared and evaluated for the presence of perceivable differences. For a product to qualify to a discrimination test, the degree and quality of the observed difference must be very small, that the two products or versions of the product can be mistakenly interchanged.

All the three selected methods were tested in a controlled laboratory condition using a sensory evaluation software called RedJade (RedJade Sensory Solutions, n.d.). The test implementation was conducted according to the strict standards of each method used (ASTM E3009-15e1, 2015; ISO 4120:2021, 2021; ISO 6658:2017, 2017; ISO 10399:2017, 2017). The samples used as stimuli were actual products of the company currently undergoing modifications. By using actual products as test samples, the test implementation was cost-effective as the results of the triangle tests were utilized to support the business decision making on whether to go ahead with the product modification or not. The three methods were tested using different product group samples both from dairy and non-dairy products. The duration of all the laboratory tests were initially planned to take approximately about two to three months, but the current restrictions held by the pandemic extended the time needed to conduct all the experiments as remote working was implemented in the company. The testing schedules were dependent on the test requests of the project groups and the availability of the products. A total of six different products were tested within the six-month duration of this experiment from the end of November 2020 to the end of May 2021.

Data analyses were executed simultaneously after each product's tests as reports were communicated to the project group within the agreed deadline. Results were automatedly analyzed by the software Red Jade. Results in each test were compared to provide a discussion and conclusion on the applicability of each three methods using different product types.

5.1 Methods

As defined in the limitations of this study, only unspecified or overall discrimination tests were compared. Among the unspecified difference testing methods, only three were

selected for comparison: triangle, constant reference duo-trio and tetrad. Each method was represented by a test series: triangle test series had three samples, duo-trio test series had a reference then two samples, and tetrad test series had four samples. All three methods were tested using the same product for comparison in a single session. All the test series were served in one tray. Each session consisted of the same product pair: control sample A and test sample B. One product pair was tested for difference using all the three methods selected for comparison. The test series included either the control sample A and its duplicate A' or the modified sample B and its duplicate B'. Both samples and the test series were randomly arranged and balanced among the participants to insure equal distribution.

In triangle test, the test series consisted of either two control (A, A') or two test samples (B, B'). The tester must select the odd sample among the three. In duo-trio test, the series consisted of a constant reference (A), one control sample (A') and one test sample (B). The tester must select the sample that matches the reference. The constant reference sample used in duo-trio test series was the current version of the product available in the market. In tetrad test, the series consisted of two control samples (A, A') and two test samples (B, B'). The tester must sort and divide the samples into two groups based on similarities. Tables 4 and 5 show the three methods selected and their test set-ups.

Table 4. Selected discrimination testing methods for comparison

Method (class of test)	Number of samples in a series	Number of possible order combination	Possible sample arrangement	Guessing probability, p
Series 1. Triangle (odddity)	3	6	AA'B, ABA', BAA', BB'A, BAB', ABB'	1/3
Series 2. Duo-trio (matching)	Ref + 2	2	Ref A+AB, Ref A+BA	1/2
Series 3. Tetrad (sorting)	4	6	AA'BB', BB'AA', ABA'B', BAB'A', ABB'A', BAA'B'	1/3

Table 5. Amount of sample servings, cups and time needed in each method

Method	Amount of sample servings needed if n=number is participants	Example if, n=20 (50 ml or g /serving)	Total amount of serving cups needed if n=20	Estimated time needed for sample and test preparation, h
Triangle	sample A= $n+1/2n$ sample B= $n+1/2n$	A=30 servings, B=30 servings	60	2
Duo-trio	sample A= $n(\text{ref})+n$ sample B= n	A=40 servings, B=20 servings	60	1
Tetrad	sample A= $n+n$ sample B= $n+n$	A=40 servings, B=40 servings	80	2,5

Duo-trio test uses the same sample number as the triangle test, but there are fewer possible sample arrangement combinations to prepare, thus making the test preparation easier, faster, less chances of making sample arrangement mistakes and more efficient. The reference sample only serving as a reminder where the two other samples are to be compared makes the discrimination task easier while inhibiting decision strategy changes among the participants.

Tetrad on the other hand has been gaining popularity as a substitute to triangle test. Many companies are considering the switch from triangle to tetrad. Several studies (Bi, 2020; J. M. Ennis, 2012; J. M. Ennis et al., 1998; Garcia et al., 2012; Ishii et al., 2014; Sanderson, 2017; Theses & Carlisle, 2014) have investigated that tetrad test requires only a third of assessors compared to triangle test as it is more sensitive in detecting small differences among samples. However, the additional fourth sample contributed to longer times of sample preparation and evaluation.

5.2 Products

The product developers and researchers in Valio R&D were informed about this project as soon as the project plan has been approved. Since the R&D sensory laboratory routinely conducts internal triangle tests with the employees as participants, the incorporation of this study with the on-going projects were both cost and time effective. The products tested

were milk drink, fruit juice, cheese, two types of yogurts and a plant-based yogurt alternative (flavored oat gurt). The schedule of each session depended on the availability of the test samples and their necessity for a discrimination test. Upon the success of the product test runs in the production plant and the availability of the products to be tested, the project group evaluated the samples first and checked if the difference in sensory quality between them was relatively small and barely noticeable. Each product was tested using all the three methods in one session at a time during the six-month duration of this experiment. The test samples (B) used were modified versions of the current products available in the market (A). The types of modification varied among the samples: some underwent process change and some had ingredient substitution, addition, or reduction in amount. In one session, the participant evaluated a total of 10 samples. Due to the number of samples evaluated at the same time, all the participants performed only a single trial of each method in one session. The list of products, the modifications made, and the sensory descriptions from the project groups is shown in Table 6. below. The sensory characteristics described below were the project group's evaluation, where some possible perceivable differences between the products could be detected by the participants.

Table 6. List of products used as stimuli.

Product category	Product type	Modification	Control sample A and A'	Test sample B and B'
Drinks, dairy	Milk	process alteration for more efficient production	neutral taste	hint of sweeter taste
Drinks, non-dairy	Fresh fruit juice	addition of an ingredient to improve nutritional value	slightly lighter in color, sourer in taste	darker in color and sweeter in taste
Spoonable snacks, dairy	Yogurt 1: Mild and one-flavor yogurt variant (simple)	reformulation and ingredient substitute to improve taste	milder taste, less full in flavor	creamier taste, fuller flavor
	Yogurt 2: Intense and mixed flavors yogurt variant (complex)	reformulation and ingredient substitute to improve taste	lighter color, milder flavor	slightly darker color, stronger flavor
Spoonable snacks, non-dairy	Fruit flavored oat-based yogurt alternative (gurt)	recipe modification to improve texture	thicker texture and mouthfeel	thinner texture and mouthfeel
Ripened hard cheese, dairy	Emmental	new starters mix from a new supplier	slightly stronger taste	milder and sweet taste

5.3 Recruitment

Once the sensory quality of the samples to be tested was confirmed by the project group, the internal recruitment for participants was launched. The recruitment was implemented using the online sensory software Red Jade, where an email invitation message was sent to all the company employees listed in the database. Internal recruitment was arranged in each product. Like all the routine recruitment organized by the R&D sensory lab, the invitation message contained all the details of the evaluation such as:

- the product to be tested and possible allergens,
- the available dates and times of testing,
- the testing location,
- the three methods to be used in the test including their short descriptions,
- the total number of samples included in the testing session, and
- the duration of the test session.

The email invitation message was written in Finnish language. The difference tests were offered for two to four consecutive days in the same week to comply with the limitations on the number of persons allowed in the sensory lab at the same time. The participants were instructed to choose their schedules for the evaluation and to participate in one session only. A thirty-minute session was reserved for each participant even though the evaluation required only ten to fifteen minutes.

5.4 Participants

The participants recruited to the tests were R&D, business, and management employees, who have prior training and experience in performing the triangle test. The degree of experience and expertise in sensory evaluation varied among the participants. None of them performed the other two methods tested in this study before.

Table 7. shows the participation of assessors in all six sessions. There were 42 different assessors recruited in total, but only five assessors have participated actively in five sessions.

Some have partaken in several sessions and some just participated once. None have participated in all six sessions. About 60 % of the total number of assessors (N=42) participated in only one or two sessions.

Table 7. Participation of assessors in all six sessions.

Number of sessions participated	Number of participants	% Participation
5	5	12 %
4	4	10 %
3	8	19 %
2	13	31 %
1	12	29 %

Each test session had different set of participants as shown in Table 8. The correct answers in each session were counted collectively. Individual performance of each participant was not evaluated separately. Recruited participants were those who reserved a testing time, while unscheduled participants were those who did not have a reserved testing time but voluntarily came to do the test. There were also cancellations on participation due to inability to come to the test in person.

Table 8. Number of assessors who participated in each session

	Milk	Juice	Cheese	Yogurt 1	Yogurt 2	Gurt
Recruited participants	17	16	20	17	18	15
Unscheduled participants	3	1	3	8	3	1
Cancelled participation due to absence	1	1		1	1	
Total participants, N	19	16	23	24	20	16

There is an existing challenge of recruiting the desired number of participants in any internal sensory tests not only in the company but in other FMCG companies as well, but the current pandemic has made this challenge more difficult to overcome. The employees were restricted on coming to work as the remote way of working was implemented, so the number of participants were lower than expected. The restrictions also did not make external testing possible for the same tests to be conducted outside the company, with non-

employees as assessors. It was initially planned to conduct at least one external test with HAMK university students as assessors, but students were also restricted to come to classes in person. Safety precautions were observed during all the test implementations. The number of assessors in the testing locations at the same time was limited to provide a safe distance between the assessors during the evaluation.

5.5 Location

The sensory evaluation took place in two centralized and controlled testing locations: the sensory lab for the R&D personnel and the testing area in the main building for the business and management personnel. The R&D sensory lab has 10 individual booths, and the main building's testing area has two. Due to the restrictions implemented in accordance with the pandemic, the participation in the testing locations were limited: only two assessors were allowed to come to the lab and only one to the main building's testing area in every 30 minutes. The participants were instructed to avoid eating or drinking strong tasting foods and drinks at least half an hour prior to their test schedules. Upon arrival to the testing location, each participant took a sample tray from the refrigerator, proceeded to the testing booth, then performed the evaluation independently using the computer with Red Jade's online questionnaire. After the evaluation, the participants were offered and allowed to take a small snack from the compensation refrigerator. The compensation snacks available varied from the company's own products like yogurts, smoothies, puddings, quarks, and ice creams.

5.6 Samples

Samples tested were current products available in the market undergoing recipe or process modification that required verification on perceptible difference. None of the tested samples were prepared or bought only for the purpose of this study. The types of products used as samples are listed in Table 9. Each product was tested using the three difference test methods in one testing session. Each participant was served with three series of samples on a tray: one series represented one difference test method. The arrangement of both the samples and the test series were randomly served and balanced among the participants. In total, each participant tasted and evaluated all the ten samples on the serving tray and

performed three difference tests in one session. As the total number of samples presented reached the maximum number of samples for discrimination testing in one session, no duplicate sessions were performed to minimize tasting fatigue.

Table 9. Number of samples to be evaluated in each difference test series

Product type	Number of samples in test series			Total number of samples in one session
	triangle	duo-trio	tetrad	
Milk	3	Ref+2	4	10
Fruit juice	3	Ref+2	4	10
Yogurt 1: simple	3	Ref+2	4	10
Yogurt 2: complex	3	Ref+2	4	10
Gurt: oat-based	3	Ref+2	4	10
Cheese	3	Ref+2	4	10

On the day of the test, samples were portioned uniformly to the number-coded cups. An example of how the samples were coded is shown in Table 10. The amount portioned to each coded cups were 50 ml for drinkable samples, 50 g for spoonable samples and two to three pieces of bite-sized cubes for cheese sample.

Table 10. An example of given number codes to milk samples in each series

MILK	TRIANGLE		DUO-TRIO		TETRAD	
Control sample, current version of the product	129	223	REF-942	524	514	845
Test sample, modified version of the product	331	413		736	687	792

The samples were arranged into three series based on the design block created by the sensory software Red Jade. An example of a design block is shown in Table 11. The triangle test series had three samples, duo-trio test series had three samples, one of which was the

reference, and tetrad test series had four samples. Both the samples and the series were randomly arranged and balanced among the participants. Each serving tray containing a total of ten samples was given a tray number to identify the corresponding sample serving arrangement. The participant code was the same as the sample tray number. The participants evaluated anonymously and identified to the test using their respective tray numbers. The blue codes were the triangle test series, the green codes were the duo-trio series, and the orange codes were the tetrad test series. Series 1 was arranged in the front row of the sample tray and the first test series to be evaluated, then Series 3 was the third row in the sample tray to be evaluated last.

The serving temperatures of each product varied on the recommended sample presentation mentioned in the company's internal sensory evaluation handbook. Milk samples were evaluated at 8-10 °C, fruit juice and cheese samples at 12-14 °C, and spoonable snacks yogurt and gurt samples at 6-8 °C. Milk and spoonable snack samples were stored and evaluated at lower temperatures compared to the suggested sample serving temperatures at 14 °C and 10 °C respectively. This precaution was done to prevent possible sample spoilage, since the samples were prepared in the morning and were stored for the whole day evaluation.

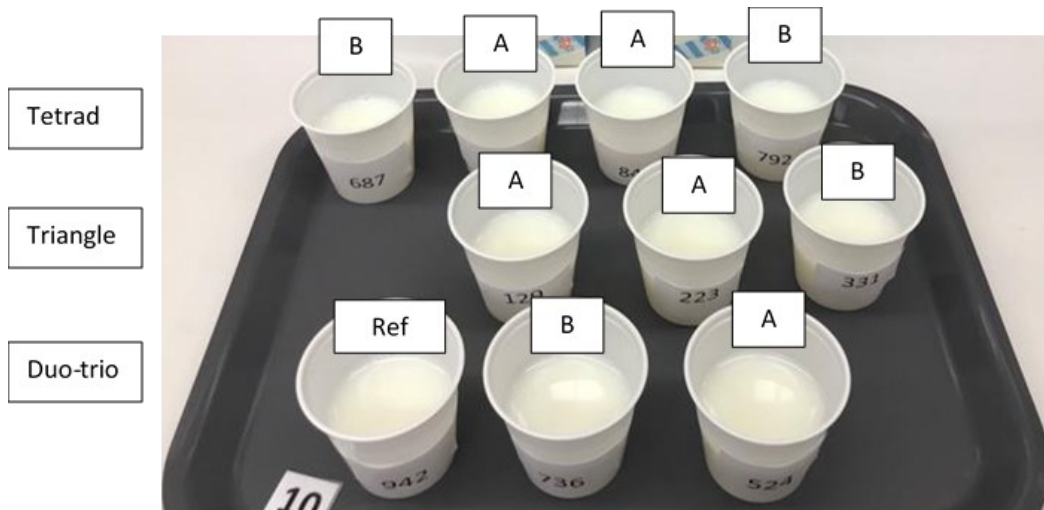
Table 11. An example of the design block showing the sample serving arrangements in each series.

DESIGN BLOCK - SAMPLE SERVING ARRANGEMENT			
Participant Code	Series1	Series2	Series3
1	129,331,413	687,792,514,845	942,736,524
2	129,331,223	942,524,736	514,845,687,792
3	942,736,524	514,687,792,845	331,129,413
4	687,514,845,792	942,524,736	129,223,331
5	687,514,792,845	331,129,223	942,736,524
6	942,524,736	331,413,129	514,687,845,792
7	942,736,524	514,687,792,845	129,331,223
8	331,129,413	942,524,736	687,792,514,845
9	331,413,129	514,687,845,792	942,524,736
10	942,736,524	129,223,331	687,514,845,792
11	514,845,687,792	129,331,413	942,736,524
12	687,514,79,845	942,524,736	331,129,223
13	331,129,223	942,524,736	514,687,792,845
14	687,514,792,845	331,129,413	942,736,524
15	942,736,524	687,792,514,845	129,331,413
16	129,223,331	942,524,736	514,845,687,792
17	687,514,845,792	129,331,223	942,524,736
18	942,736,524	514,687,845,792	331,413,129

An example of a sample tray is shown in Figure 16. Tray number 10 had duo-trio test as the first series and tetrad test was the last series. Sample A is the current version and sample B is the modified version. The picture shows how the samples are randomly distributed in each series. Codes A and B also reveals the correct answers in each series: the participant must

choose sample 524 as the sample that matches the reference (duo-trio), sample 331 as the odd sample among the three (triangle) and sorted samples 687 and 792 together in one group (tetrad). The participants did not have any prior information which sample codes were the reference (A) and which ones were the test sample (B). Pictures of the sample trays presented in each product were included in Appendix 6.

Figure 16. An example of the milk's sample tray.



5.7 Test questionnaire

The evaluations were performed independently by the participants. On the first page of the questionnaire, the participants were informed about discrimination test in general and what are the other methods available aside from triangle test. Then, the second page described the three methods to be compared and performed during the session and how do they differ from each other. This allowed them to prepare for the actual evaluation. After the information pages, the three difference test series followed and were performed at random orders by the participants. In between the test series was a mandatory one-minute break for palate cleanse with water and in some cases with unsalted crackers as well. Table 12 describes the instructions in each method, but the exact questionnaire used throughout the experiment is included in the Appendix section.

Table 12. Steps in the evaluation part

Questionnaire part	Description
Info page	Definition of discrimination tests in general
	Description of the three selected methods to be performed during the session
Difference tests	<u>Triangle:</u> <i>Choose the most different/odd sample. (Oddity)</i>
	<u>Constant reference duo-trio:</u> <i>Match sample to the reference. (Matching)</i>
	<u>Tetrad:</u> <i>Sort into two groups based on similarities. (Sorting)</i>
Degree of difference	<i>State the degree of difference (DoD) between the samples using the scale:</i> 1=no difference 2=small, just noticeable difference 3=clear difference, 4=very clear difference <i>If DoD >0, How are the samples different? Open comments</i>
Method comparison	Duo-trio/Tetrad vs. Triangle <i>What do you think about performing duo-trio/tetrad test compared to the triangle test?</i> 1=a lot easier than triangle (+ open comments), 2=slightly easier than triangle (+ open comments), 3=as easy as/as difficult as, 4=slightly more difficult than triangle (+ open comments), 5=a lot more difficult than triangle (+ open comments)
Background	Gender and age group
Feedback	Revealing the correct answers for training purposes + optional retasting, then participant's overall feedback (optional)

As an extension of the basic difference test, a four-point scale degree of difference question was included to give the project group more information on the nature of the products' differences. The scale used were 0=no difference, 1=small, just noticeable difference, 2=clear difference, and 3=very clear difference. Those who noticed the difference were asked to describe how the samples differ from one another. This extension of difference testing provided important information to the project group especially when the samples were perceived as different. The extended difference test for all the three methods were

performed by participants with milk and juice samples. For the rest of the samples, the degree of difference question was asked only in triangle test to minimize the evaluation tasks of the participants.

After the three difference test series were performed, participants were then asked to rate the difficulty or easiness of performing the two other methods compared to triangle test. A five-point scale was used for method comparison: 1=a lot easier than triangle test, 2=slightly easier than the triangle test, 3=as easy or as hard as the triangle test, 4=slightly more difficult than the triangle test, and 5=a lot more difficult than the triangle test. The sample codes and the correct answers to each series were revealed to the participants after the test session to allow them to check their answers. Revealing the answers to the participants after the test and allowing them to practice with their samples again was found to be useful as a part of individual training for detection of small differences. This practice allows the assessors to develop and improve the evaluation performance in the future. Personal background details were also gathered such as gender and age group. At the end, they were also allowed to give optional feedback about the products or the test itself.

5.8 Data collection

Data was gathered simultaneously in RedJade as the test progressed. Live results were available to the test organizers through the sensory software, making the evaluation process easy to follow. Table 13 is an example of the data collected along with the topline analysis results.

Table 13. An example of a data sheet with topline results

Participants	MILK		
	Triangle	Duo trio	Tetrad
1	✓	X	X
2	X	✓	✓
3	X	✓	✓
4	X	✓	X
5	✓	X	X
6	X	✓	✓
7	X	X	X
8	X	X	X
9	X	X	X
10	✓	✓	X
11	✓	X	✓
12	X	X	X
13	✓	X	✓
14	X	✓	X
15	X	X	X
16	X	X	X
17	X	✓	✓
18	X	X	✓
19	X	✓	X
Minimum correct required	11	14	11
Total Correct	5	8	7
Total Responses	19	19	19
Percent Correct	26 %	42 %	37 %
Confidence Level (One-Tailed)	19 %	18 %	54 %
P-Value / Alpha Risk (One-Tailed)	0.8121	0.8204	0.4569

5.9 Data analysis

As described in the methods description earlier, the number of correct responses were counted in total. The individual scores were not directly of interest. Those who answered incorrectly were not counted. The minimum numbers of correct responses to reject the null hypothesis of 'no difference' at the selected significance level according to the total number of assessors, 'n' in each method were provided in Table 14. The values in the table were derived from the sensory methodology standards (ASTM E3009-15e1, 2015; ISO 4120:2021, 2021; ISO 10399:2017, 2017). If the total number of correct responses is equal to or greater than the minimum number of correct responses required to establish significant difference, the null hypothesis that the two products are perceived as similar will be rejected. Since

statistical tables were made available, results can easily be interpreted even without using a separate statistical software. Using a sensory software like Red Jade, provided statistical analysis to obtain the exact p -values for comparison. As defined prior to the test, the significance level of $p = 0,05$ was set, which was a common standard used in food industry. With the computed values of p , the results were also compared to match the results from the statistical significance tables. If p -value is equal or less than 0,05 the perceived difference was statistically significant but when p -value is greater than 0,05 the perceived difference between the two samples was not significant. Additional data analyses were performed with Microsoft Excel to explore the different variable relations.

Table 14. Minimum number of correct responses needed to establish significance at probability level of 5 % for duo-trio, triangle, and tetrad tests

n	$\alpha = 0,05$		
	triangle	duo-trio	tetrad
3	-	-	3
4	-	-	4
5	4	-	4
6	5	6	5
7	5	7	5
8	6	7	6
9	6	8	6
10	7	9	7
11	7	9	7
12	8	10	8
13	8	10	8
14	9	11	9
15	9	12	9
16	9	12	9
17	10	13	10
18	10	13	10
19	11	14	11
20	11	15	11
21	12	15	12
22	12	15	12
23	12	16	12
24	13	17	13

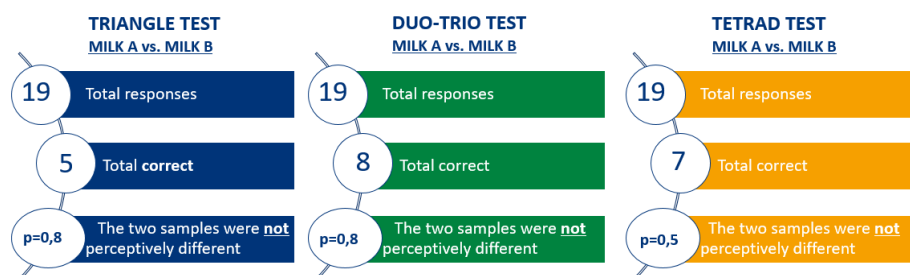
6 Results and discussion

Only the results of the three difference tests and the methods comparison will be presented and discussed in this section. The results from the degree of difference questions were presented in the Appendix section, as it was not included in all the three methods tested using all the products.

6.1 Milk

Figure 17 shows that among 19 assessors only 5, 8 and 7 answered correctly in triangle, duo-trio, and tetrad tests respectively. All the methods produced the same results with a p-value of greater than 0,05. Reference milk A and test milk B do not have a statistically significant difference in sensory qualities using all the three methods. Referring to the statistical table, the minimum number of correct answers needed to establish significant difference in triangle and in tetrad tests is 11 and in duo-trio test is 14, if the panel size is 19 (N=19). Since the minimum number of correct responses was not met in all the three methods, there was no significant evidence that sensory difference exists between the two milk samples. Therefore, the process change was not perceivable.

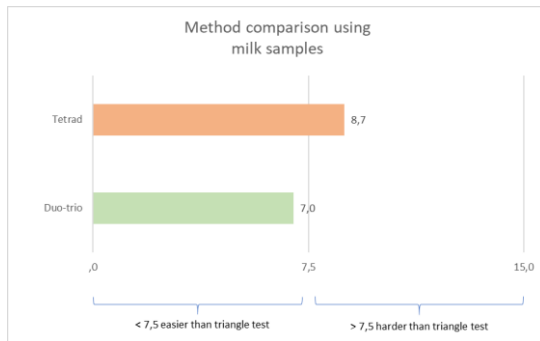
Figure 17. Results of discrimination tests with milk samples



The methods comparison rating in milk was different from the rest of the products (Figure 18). The scores were rated in a 15-point line scale: scores less than 7,5 = easier than triangle test, score 7,5 = as easy or as difficult as the triangle test, and scores greater than 7,5 = more difficult than the triangle test. With milk samples (n=19), duo-trio was rated slightly easier (7,0) than triangle, and tetrad was slightly harder (8,7) compared to triangle test. Based on

the feedbacks given by the participants, the scale was difficult to use as the two methods were new to them, so the scale was simplified and narrowed down to a five-point scale for the rest of the sessions.

Figure 18. Method comparison's results - Milk



Assessors were also asked to explain their answers if they chose easier or more difficult than triangle test as shown in Table 15. Those who rated duo-trio as easier than triangle (N=4) commented that the reference sample was helpful, and it was easier to focus on choosing the correct answer between two samples. However, some said that duo-trio was harder to perform than the triangle test (N=3) due to the small differences between the two samples. The new method's instructions where samples were to be compared to a reference sample was difficult. Three assessors rated tetrad as easier than triangle with a comment that pairing the samples into two groups gave a confirmation on the choice of difference rather than choosing just one odd sample in triangle test. However, more assessors rated tetrad as more difficult to perform than triangle test (N=7) due to having more samples to evaluate.

Table 15. Method comparison's open comments - Milk

<p>Duo-trio easier than triangle, N=4 (7,1)</p> <ul style="list-style-type: none"> • vertailunäyte on tiedossa • On paljon helpompaa kuin tietää mikä näytteistä on jonkun pari. • Osaa etsiä oikeita asioita näytteestä. Esim juustossa riippuen palasta voi olla muitakin eroja kuin vain se kysytty. Ei keskitytä vähäpätöiseen eroon josta ei ole kehityksessä hyötyä ja menee esim normaalin vaihtelun piiriin • Siinä pitää valita vain kahdesta vaihtoehdosta. Tällöin vastaus menee 50% todennäköisyydellä oikein. 	<p>Tetrad easier than triangle, N=3 (7,0)</p> <ul style="list-style-type: none"> • na • Useampi näyte antaa vahvistusta sille, että etsii oikeanlaista eroa. • Kalmitestissä on löytävinnään jonkun eron, mutta sitä ei voi vahvistaa rinnakkaisesta näytteestä. • ei ollut helpompi
<p>Duo-trio harder than triangle, N=3 (10,8)</p> <ul style="list-style-type: none"> • kun erot on pieniä niin aisti ei enää toimi usean näytteen kesken niin hyvin kuin kolmitestissä • Tuotteet pitää maistaa järjestyksessä vasemmalta oikealle, ei voi vertailla keskenään eri näytteitä. • Minun oli vaikeampi löytää eroa verrattuna vertailuun 	<p>Tetrad harder than triangle, N=7 (11,1)</p> <ul style="list-style-type: none"> • kun erot on pieniä niin aisti ei enää toimi usean näytteen kesken niin hyvin kuin kolmitestissä • Tuotteet pitää maistaa järjestyksessä vasemmalta oikealle, ei voi vertailla keskenään eri näytteitä. • enemmän vaihtoehtoja, enemmän maistettavaa • 1 näyte enemmän kuin kolmitesti • enemmän maistettavia näytteitä, varsinkin kun en oikeastaan havainnut niissä mitään eroa • näytteitä oli enemmän arvioitavana ja kaksi oli samanlaisia • niin paljon maistettavaa

6.2 Juice

Results in Figure 19 shows that among 16 assessors 12, 10 and 11 answered correctly in triangle, duo-trio, and tetrad respectively. The minimum number of correct answers needed to establish significant difference in triangle and in tetrad tests is 9 and in duo-trio test is 12, if the panel size is 16 (N=16). The minimum number of correct answers was not met only in duo-trio test. Reference juice A and test juice B do have a statistically significant difference in sensory qualities using triangle and tetrad methods. However, duo-trio test did not produce the same results. Since there was significant evidence that sensory difference exists between juice A and juice B using the triangle and tetrad tests, the addition of a value-added ingredient was perceivable. The conclusion communicated to the project group that perceptible difference exists was based on the current method used in the company, which is the triangle test.

Figure 19. Results of discrimination tests with juice samples

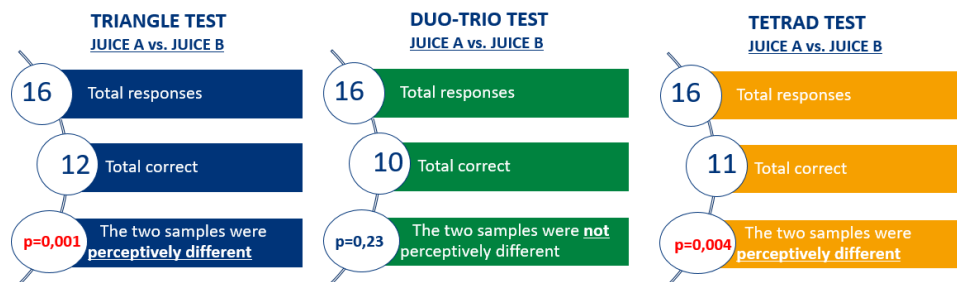
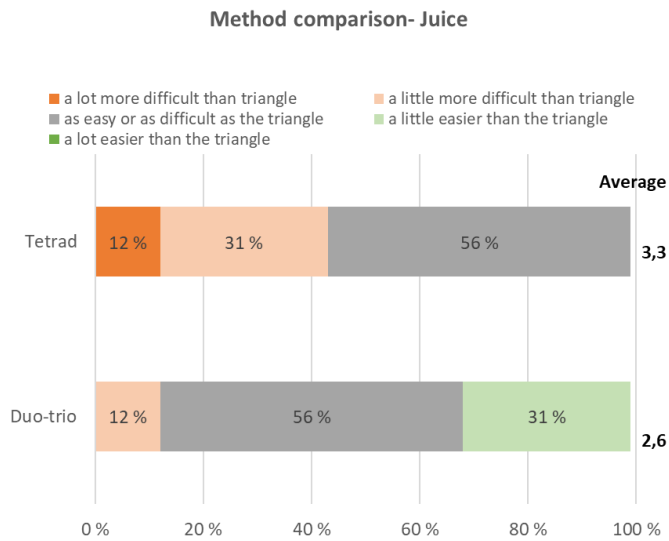




Figure 20 shows the method comparison results in juice (N=16). In here, the scale was changed from a 15-point line scale to a five-point scale. This scale was used throughout the rest of the products in this experiment. The reason why the scale was changed was due to the feedback from the assessors. Using only a five-point scale, the comparison was simpler and the scale was easier to use. Scores from 1-2 were rated as easier, 3 was as easy or as difficult and 4-5 points were given if the method is more difficult than the triangle test. Duo-trio was rated slightly easier (2,8) and tetrad was slightly harder (3,6) compared to triangle test. Although half of the assessors rated duo-trio as easy/as hard as the triangle test, a third said the method was easier but the rest said it was harder. In tetrad, about half also said the method is as easy/as hard, but the other half said it was harder than the triangle test.

Figure 20. Method comparison's results - Juice



As seen in Table 16, for five testers duo-trio was easier due to the availability of the reference sample. Two persons said it was harder than the triangle test but failed to justify their scores. Nobody scored tetrad as easier than triangle but seven assessors rated tetrad as more difficult due to more samples needed to evaluate.

Table 16. Method comparison's open comments - Juice

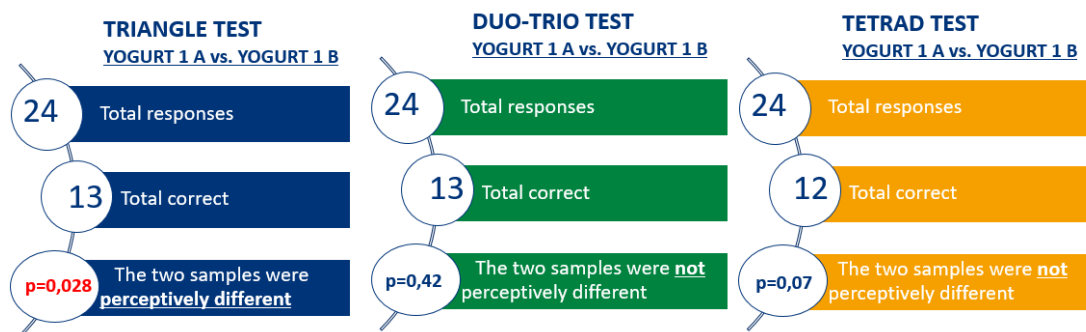
	Duo-trio easier than triangle 2p. (N=5) <ul style="list-style-type: none"> tietää mihin verrataan on näyte johon verrata On jokin referenssi mihin verrata, helpottaa arviointia vertailunäyte ja sen parin löytäminen Kun tiesit, että mikä näyte on toinen parista, erilainen tuntui helpommalta arvioida. 	Tetrad easier than triangle (N=0)
	Duo-trio harder than triangle 4p. (N=2) <ul style="list-style-type: none"> n/a ei eroa 	Tetrad harder than triangle 4,3p. (N=7) <ul style="list-style-type: none"> n/a pareihin jakaminen tuntui vaikeammalta kuin yhden eroavan löytäminen enemmän näytteitä Useampi näyte joutuu maistamaan yhden näytteen enemmän ja se rasittaa makuuistia Vaihtoehtoja oli enemmän, siksi vaikeampi. pitää maistella enemmän näytteitä

6.3 Yogurt

Yogurt 1: mild and one-flavored variant (simple)

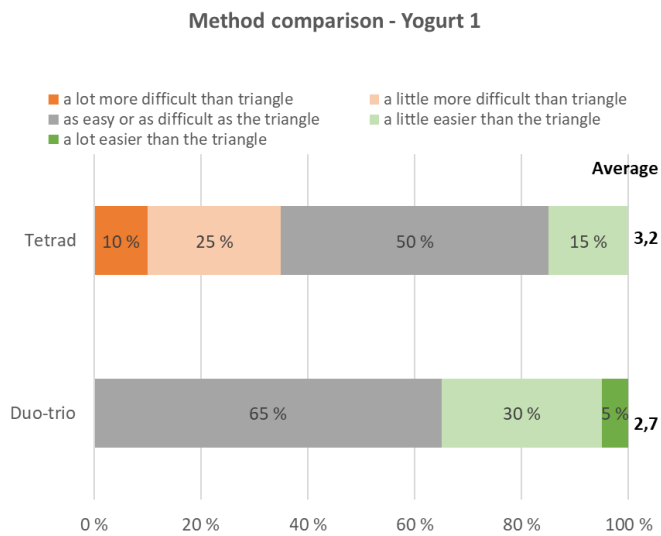
Results in Figure 21 shows that among 24 assessors, almost half have answered correctly in all the tests, but only triangle test resulted with a p-value $< 0,05$. The minimum number of correct answers needed to establish significant difference in triangle and in tetrad tests is 13 and in duo-trio test is 17, if the panel size is 24 (N=24). Reference yogurt 1 A and test yogurt 1 B do have a statistically significant difference in sensory qualities using triangle test method ($p < 0,05$). Both the tetrad and the duo-trio tests did not show an evidence of sensory difference ($p > 0,05$). As a conclusion, there was a statistically significant evidence that difference exists between Yogurt 1 A and Yogurt 1 B in triangle test. In the contrary, tetrad and duo-trio tests' correct answers were not enough to conclude that the sensory difference was noticeable. Since triangle test is the current method of use, the conclusion was made based on its results.

Figure 21. Results of discrimination tests with yogurt 1 samples



The method comparison in Figure 22 shows that using yogurt 1 (N=24), duo-trio was rated slightly easier (2,7) than triangle and tetrad was slightly harder (3,2) compared to triangle test. About one third of the assessors rated duo trio as easier and the rest rated as easy/as difficult as the triangle. With yogurt 1 as samples, tetrad versus triangle showed a more polarized results with some rated tetrad as harder but some said it was easier. While tetrad was more confusing than the duo-trio, almost half of the participants rated the method as the same as triangle test. No one rated duo-trio test as more difficult than triangle test.

Figure 22. Method comparison's results - Yogurt 1



When asked about their explanations summarized in Table 17, duo-trio being rated easier than triangle (N=7) was due to having only two test samples to choose from. The existence of reference sample was also an advantage. In tetrad, four testers said it was easier to group the samples into pairs. One also commented that tetrad was easier because it was the first series to be evaluated. To some testers (N=6), tetrad is more difficult because there were more samples to evaluate making it harder to decide which ones to group together.

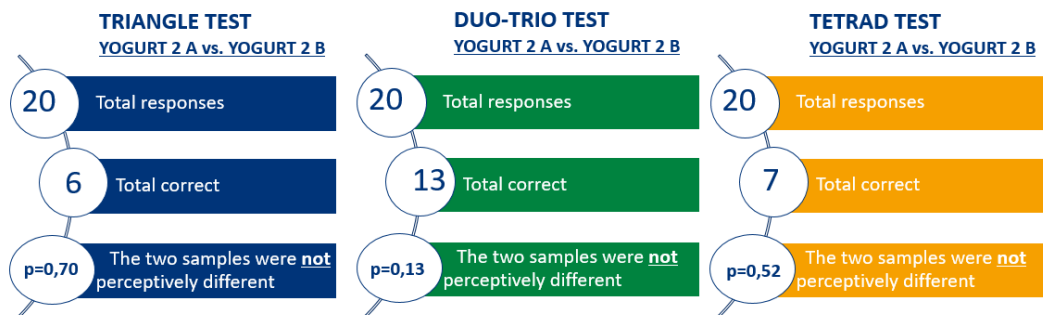
Table 17. Method comparison's open comments - Yogurt 1

	Duo-trio easier than triangle 1,9p. (N=7) <ul style="list-style-type: none"> Oli helpompi löytää pari kuin yksi kolmesta vain kaksi vaihtoehtoa vertailunäyte auttaa arvioinnissa. on verrokki johon verrata on näyte mihin voi verrata kolmitestissä monesti huomaa eron kahden ekan näytteen välillä, mutta kolmannelta ei keksi kumpaa ensimmäisistä se muistuttaa en osaa selittää, miksi oli helpompaa 	Tetrad easier than triangle 2p. (N=4) <ul style="list-style-type: none"> Kahteen ryhmään jako tuntui helpommalta Helpompi löytää erot, kun oli enemmän näytteitä saattoi johtua siitä että oli ekana ja suu vielä 'tuore' pareja on helpompi etsiä kuin paritonta
	Duo-trio harder than triangle (N=0)	Tetrad harder than triangle 4,3p. (N=6) <ul style="list-style-type: none"> enemmän näytteitä vaikeampi jakaa kahteen joukkoon 4kpl enemmän maistettavaa, tuntuu enemmän vaikea hallita niin montaa näytettä kolme näytettä helpompi kuin neljä, kun erot pieniä. enemmän näytteitä vaikeampi erottaa eroja

Yogurt 2: intense and mixed-flavored variant (complex)

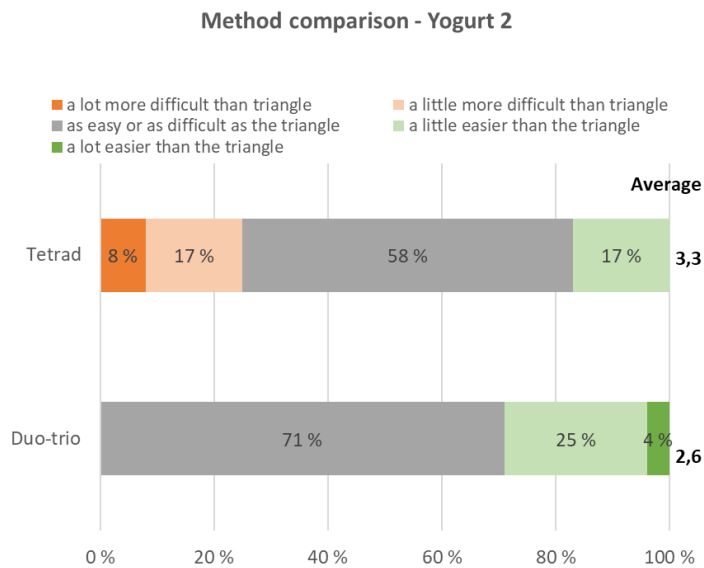
In Figure 23, among 20 assessors 6, 13 and 7 answered correctly in triangle, duo-trio, and tetrad respectively. The minimum number of correct answers needed to establish significant difference in triangle and in tetrad tests is 11 and in duo-trio test is 15, if the panel size is 20 (N=20). The minimum number of correct answers was not met in in all three methods. Reference yogurt 2 A and test yogurt 2 B do not have a statistically significant difference in sensory qualities using all the three methods ($p > 0,05$). As a conclusion, since there was no significant evidence that sensory difference exists between yogurt 2 A and yogurt 2 B in all the tests, the product reformulation was not perceivable.

Figure 23. Results of discrimination tests with yogurt 2 samples



In Figure 24, using yogurt 2 as samples (N=20) duo-trio was rated slightly easier (2,6) than triangle and tetrad was slightly harder (3,3) compared to triangle test. As observed in yogurt 1, tetrad was also more polarized compared to duo-trio test. There were testers who rated tetrad as more difficult and to some tetrad was easier. Duo-trio on the other hand, was often rated as easier than triangle.

Figure 24. Method comparison's results - Yogurt 2



A summary of open comments in Table 18 shows that duo-trio was rated easier (N=7) due to the help of the reference sample. With tetrad, it was both rated easier (N=3) because of the seemingly easier task to choose similar samples and more difficult (N=7) due to more samples needed to evaluate.

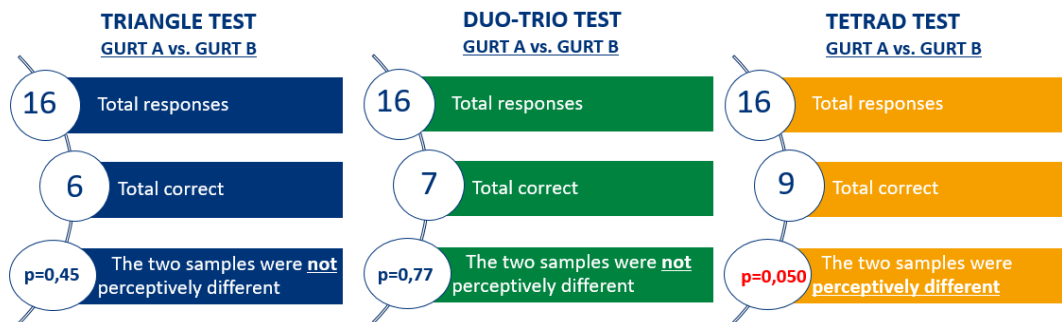
Table 18. Method comparison's open comments - Yogurt 2

	Duo-trio easier than triangle 1,9p. (N=7)	Tetrad easier than triangle 2p. (N=3)
	<ul style="list-style-type: none"> • <i>Helpompi löytää pari kuin yksi kolmesta</i> • <i>voi verrata referenssiin</i> • <i>koska oli verrokki</i> • <i>on malli</i> • <i>en osaa selittää</i> • <i>Kolmitestissä vaikea kun samanaikaisesti pitää etsiä samanlaista ja erilaista</i> • <i>vertailunäyte poistuu arvioitavista näytteistä</i> 	<ul style="list-style-type: none"> • <i>Helpompi tunnistaa ero, kun oli useampia näytteitä</i> • <i>jotenkin helpompi tunnistaa kaksi samaa vs. yksi poikkeava</i> • <i>parien etsiminen tuntuu helpommalta</i>
	Duo-trio harder than triangle (N=0)	Tetrad harder than triangle 4,3p. (N=7)
		<ul style="list-style-type: none"> • <i>enemmän näytteitä maistettavana</i> • <i>jakaminen kahteen ryhmään, enemmän näytteitä</i> • <i>vaikea löytää pareja, jostain syystä kolmen seasta on helpompi löytää poikkeava.</i> • <i>liikaa vaihtoehtoja</i> • <i>joutuu maistamaan enemmän</i> • <i>no kun on niin monta näytettä eikä viimeisen kohdalla ainakaan muista miltä ensimmäinen maistui</i> • <i>näin pienellä erolla, aistit menevät vain sekaisin jos on enemmän näytettä</i>

6.4 Plant-based yogurt alternative (*gurt*)

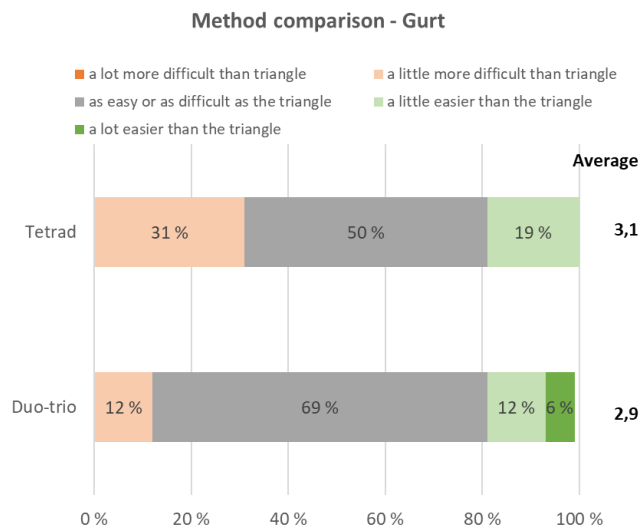
Results in Figure 25 shows that among 16 assessors 6, 7 and 9 answered correctly in triangle, duo-trio, and tetrad tests respectively. The minimum number of correct answers needed to establish significant difference in triangle and in tetrad tests is 9 and in duo-trio test is 12, if the panel size is 16 (N=16). The minimum number of correct answers was met only in tetrad test. Reference gurt A and test gurt B do not have a statistically significant difference in sensory qualities using triangle and duo-trio tests ($p>0,05$). With tetrad test, sensory difference was perceived ($p=0,05$). As a conclusion, since there was no significant evidence that sensory difference exists between gurt A and gurt B ($p>0,05$) in the current method triangle and also in duo-trio, the reduction in amount of an ingredient was not perceivable.

Figure 25. Results of discrimination tests with gurt samples





The method comparison results in Figure 26 shows that with plant-based yogurt alternative as samples (N=16) duo-trio was rated slightly easier (2,9) and tetrad was rated slightly harder (3,1) compared to triangle test. Both tetrad and duo-trio tests were rated easier as well as more difficult than the triangle test. Almost one third of the participants rated tetrad as more difficult than triangle, but duo-trio was harder only to a tenth of the testers. Both methods were rated easier by one fifth of the total participants.

Figure 26. Method comparison's results – Gurt



Open comments in Table 19 shows why the methods were rated easier or harder than triangle test. Three testers rated duo-trio as easier than triangle due to the easier task of comparing to the reference sample while two testers said it was difficult to perform due to being used to triangle test and comparing to the reference sample as a task being hard. In tetrad, three testers said it was easier to group the samples into pairs while other three testers said it was difficult because there were more samples to taste.

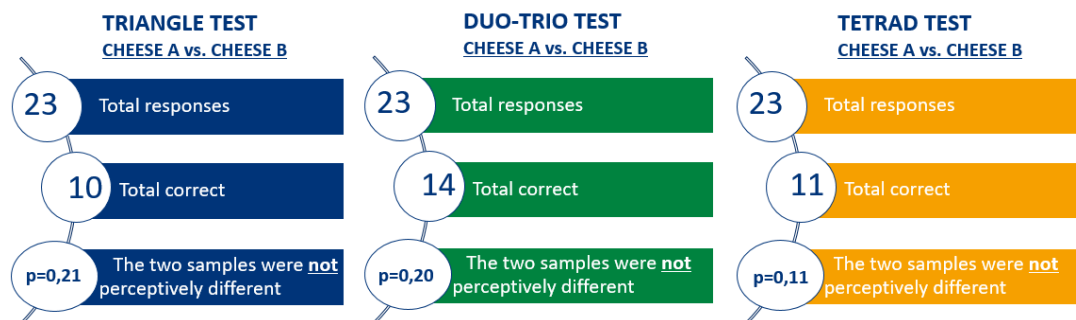
Table 19. Method comparison's open comments - Gurt

	Duo-trio easier than triangle 1,7p. (N=3) <ul style="list-style-type: none"> tietää mitä etsii Helpompia verrata, kun tietää, mihin vertaa Referenssiin vertaaminen auttaa tekemään päätöksen 	Tetrad easier than triangle 2p. (N=3) <ul style="list-style-type: none"> pareja helpompi etsiä Enemmän vaihtoehtoja selkeä ja helppo, jotenkin tuntui helpommalta tehdä vertailua ryhmittelemällä
	Duo-trio harder than triangle 4p. (N=2) <ul style="list-style-type: none"> tottunut kolmitestiin vaikeampi vertailla 	Tetrad harder than triangle 4p. (N=3) <ul style="list-style-type: none"> ei muista miltä eka maistui kun pääsee vikaan paljon maisteltavaa enemmän vaihtoehtoja ja maistettavaa näytteitä niin paljon että makuaisi turtuu Enemmän näytteitä arvioitavana

6.5 Cheese

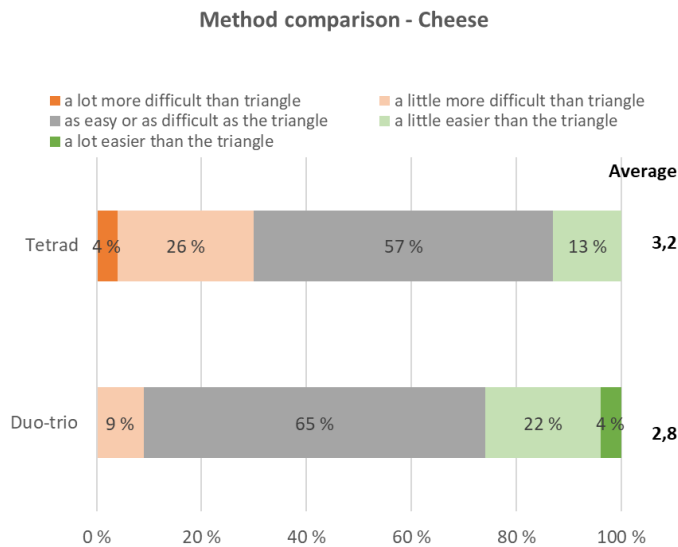
Results in Figure 27 shows that among 23 assessors 10, 14 and 11 answered correctly in triangle, duo-trio, and tetrad respectively. The minimum number of correct answers needed to establish significant difference in triangle and in tetrad tests is 12 and in duo-trio test is 16, if the panel size is 23 (N=23). The minimum number of correct answers was not met in all three methods. Reference cheese A and test cheese B do not have a statistically significant difference in sensory qualities using all the three methods ($p > 0,05$). In conclusion, since there was no significant evidence that sensory difference exists between cheese A and cheese B in all tests, the replacement of an ingredient from a new supplier was not perceivable.

Figure 27. Results of discrimination tests with cheese samples





In method comparison using cheese samples (N=23) duo-trio was rated slightly easier (2,8) than triangle and tetrad was slightly harder (3,2) compared to triangle test. Although at least half of the testers rated both methods being as easy/as hard as the triangle, almost one in every four testers rated duo-trio being easier, while a third rated tetrad being harder to perform.

Figure 28. Method comparison's results - Cheese



The opinion of the testers regarding the degree of difficulty of both methods compared to triangle test was divided. Six testers told that duo-trio was easier because of the reference sample but six testers also told that tetrad was harder due to the number of samples to be tasted. There were comments on why duo-trio was harder (N=2) but the explanations were confusing as both were neither related to the method being rated nor the scores given for comparison. Tetrad test while involving an additional sample was still rated easier than triangle (N=3) because the task of searching for two pairs seemed to be easier.

Table 20. Method comparison's open comments - Cheese

	Duo-trio easier than triangle 1,8p. (N=6) <ul style="list-style-type: none"> • vertailu auttaa • Tietää mitä etsiä • Voi suoraan verrata referenssiin • vertailunäyte mukava, vähemmän testinäytettä • On varma, että toinen on samanlainen • Helpompaa kun tietää mihin vertailla 	Tetrad easier than triangle 2p. (N=3) <ul style="list-style-type: none"> • jos näytteiden välillä on eroa, on suhteellisen helppoa yhdistää 2 näytettä toisiinsa • helpompi maistaa eroja kun voi maistaa kahdesti • Kahden näytteen ulkonäkö. Pinnalla kosteutta
	Duo-trio harder than triangle 4p. (N=2) <ul style="list-style-type: none"> • näytteitä enemmän • erot oli selkeämmät 	Tetrad harder than triangle 4,1p. (N=6) <ul style="list-style-type: none"> • paljon maisteltavaa • NA • pitää maistaa useampaa näytettä • Jos ei ole varma missä ominaisuudessa ero ja jokainen näyte poikkeaa jollain tavalla toisesta • enemmän näytteitä ja enemmän vertailtavaa • Enemmän arvioitavia juustoja.

6.6 Summary

As the null hypothesis that the products are close enough to be perceived as similar, the chance that the assessors will return a correct answer is equal to the guessing probability of the method used for evaluation. If the probability of gaining the correct responses is greater than the guessing probability, the null hypothesis will then be rejected and the products will be concluded to be perceptibly different. The results of three difference test methods performed simultaneously using different products are summarized in Table 21.

Table 21. Results of the three difference tests methods

Product	Method	Null hypothesis, H_0	Alternative hypothesis, H_a	Proportion correct, P_c	Minimum correct answers needed, x	No significant difference (NSD) <i>p-value</i> > 0,05
Milk	Triangle	$p=1/3$	$p > 1/3$	5/19 correct	11	NSD
	Duo trio	$p=1/2$	$p > 1/2$	8/19 correct	14	NSD
	Tetrad	$p=1/3$	$p > 1/3$	7/19 correct	11	NSD
Juice*	Triangle	$p=1/3$	$p > 1/3$	12/16 correct	9	$p = 0,0008$
	Duo trio	$p=1/2$	$p > 1/2$	10/16 correct	12	NSD
	Tetrad	$p=1/3$	$p > 1/3$	11/16 correct	9	$p = 0,0040$
Cheese	Triangle	$p=1/3$	$p > 1/3$	10/23 correct	12	NSD
	Duo trio	$p=1/2$	$p > 1/2$	14/23 correct	16	NSD
	Tetrad	$p=1/3$	$p > 1/3$	11/23 correct	12	NSD
Yogurt 1*	Triangle	$p=1/3$	$p > 1/3$	13/24 correct	13	$p = 0,0284$
	Duo trio	$p=1/2$	$p > 1/2$	13/24 correct	17	NSD
	Tetrad	$p=1/3$	$p > 1/3$	12/24 correct	13	NSD
Yogurt 2	Triangle	$p=1/3$	$p > 1/3$	6/20 correct	11	NSD
	Duo trio	$p=1/2$	$p > 1/2$	13/20 correct	15	NSD
	Tetrad	$p=1/3$	$p > 1/3$	7/20 correct	11	NSD
Gurt*	Triangle	$p=1/3$	$p > 1/3$	6/16 correct	9	NSD
	Duo trio	$p=1/2$	$p > 1/2$	7/16 correct	12	NSD
	Tetrad	$p=1/3$	$p > 1/3$	9/16 correct	9	$p = 0,0500$

All the three methods produced the same results of no significant difference (NSD) in products like milk, cheese, and mixed-flavored yogurt 2. On the contrary, a significant difference was observed only from triangle test in mild-flavored yogurt 1 samples. A significant difference result was also observed only from tetrad test in gurt samples. Meanwhile, both triangle and tetrad methods produced a significant difference result in juice samples. The duo-trio tests failed to produce significant difference results in juice,

yogurt 1, and gurt. The sensitivity of each method in detecting perceivable differences based on the number of correct responses observed cannot be compared directly due to the higher probability of guessing in duo-trio test. Between triangle and tetrad test methods, although the difference in number of correct answers between the two was small, tetrad produced more correct answer in most of the products tested.

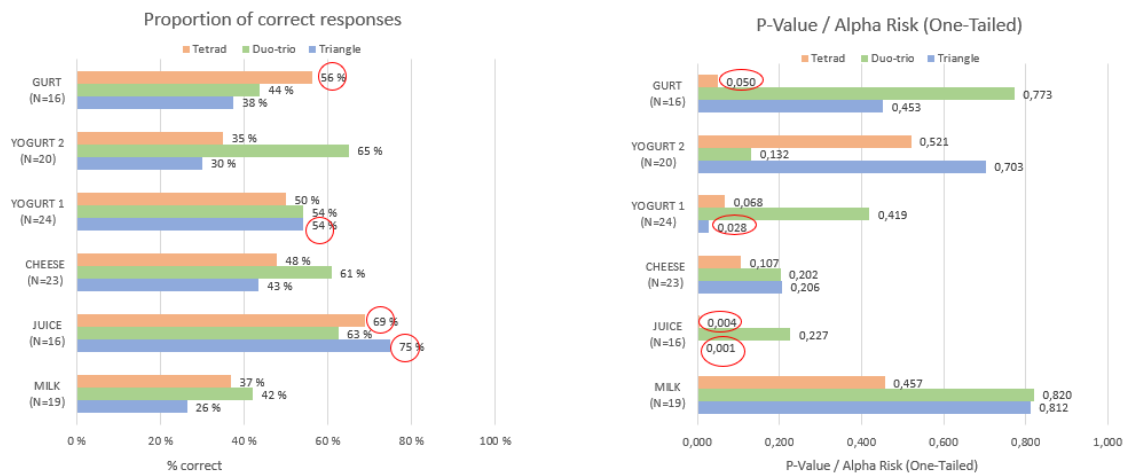
Table 22. Summary of percent correct answers, p-values, and confidence levels

Sample	Method	Guessing probability, p	Min. correct, x	Total Responses, n	Total Correct, Tc	Percent correct, Pc	p-value/Alpha risk (One-Tailed)	Confidence Level (One-Tailed)
Milk	Triangle	33 %	11	19	5	26 %	0,81	19 %
	Duo trio	50 %	14	19	8	42 %	0,82	18 %
	Tetrad	33 %	11	19	7	37 %	0,46	54 %
Juice*	Triangle	33 %	9	16	12	75 %	0,00	100 %
	Duo trio	50 %	12	16	10	63 %	0,23	77 %
	Tetrad	33 %	9	16	11	69 %	0,00	100 %
Cheese	Triangle	33 %	12	23	10	43 %	0,21	79 %
	Duo trio	50 %	16	23	14	61 %	0,20	80 %
	Tetrad	33 %	12	23	11	48 %	0,11	89 %
Yogurt 1*	Triangle	33 %	13	24	13	54 %	0,03	97 %
	Duo trio	50 %	17	24	13	54 %	0,42	58 %
	Tetrad	33 %	13	24	12	50 %	0,07	93 %
Yogurt 2	Triangle	33 %	11	20	6	30 %	0,70	30 %
	Duo trio	50 %	15	20	13	65 %	0,13	87 %
	Tetrad	33 %	11	20	7	35 %	0,52	48 %
Gurt*	Triangle	33 %	9	16	6	38 %	0,45	55 %
	Duo trio	50 %	12	16	7	44 %	0,77	23 %
	Tetrad	33 %	9	16	9	56 %	0,05	95 %

The values of the minimum number of correct answers required to establish significant difference at the predetermined probability level of 5 % were obtained from the standards of each method. If the total number of correct responses is higher than the minimum number of correct responses required to establish significant difference ($Tc \geq x$), sensory difference is perceived. If the computed p-value is equal to or less than 0,05 ($p\text{-value} \leq 0,05$), the sensory difference is statistically significant. As the probability level was predetermined before the test, so as the confidence level of 95 % was allowed. The exact values of the confidence level were also computed by the sensory software. The results of juice, yogurt 1 and gurt, where difference was perceived ($p < 0,05$) were statistically significant at a confidence level of at least 95 %.

An illustration of all the percent correct values in each method with the computed p-values is illustrated in Figure 29. In gurt samples, only the tetrad test gathered the most correct answers that resulted in a p-value =0,05. In yogurt 1, the triangle test gathered the greatest number of correct answers reaching the p-value of < 0,05. Duo-trio test in yogurt 1 also had the same percent correct value but the method's probability of guessing the correct answer by chance is 1/2. Therefore, there are more correct answers required for duo-trio test to meet the same significance level as the triangle test. For the juice samples, both triangle and tetrad reached high percent correct answers which were enough to establish significant difference between samples at p-values of < 0,05.

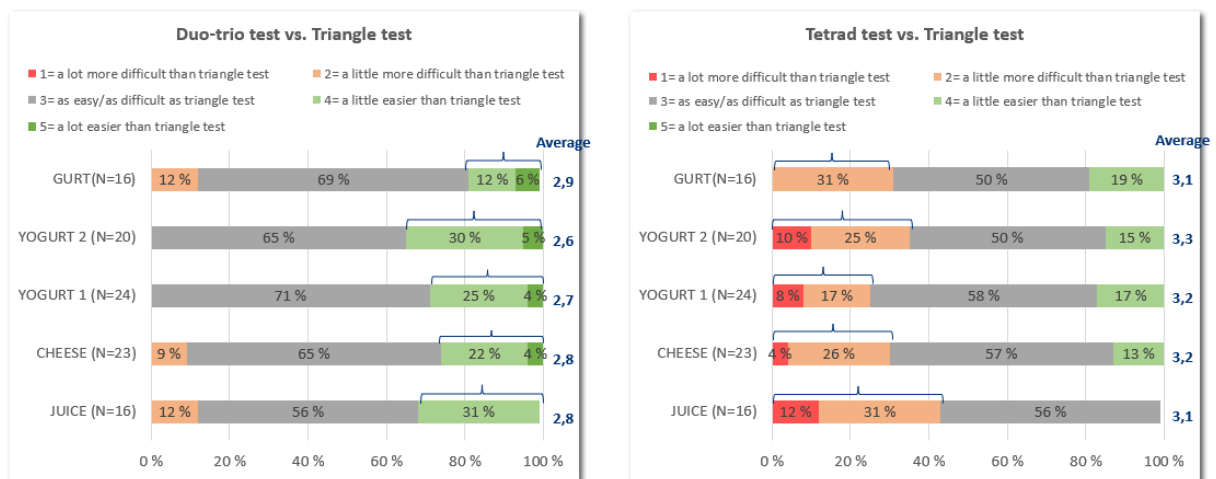
Figure 29. Summary of all the percent correct answers and computed p-values



All the methods with the encircled p-values of equal to or less than 0,05 were considered to establish an evidence that the difference exists between two samples. In difference testing, the null hypothesis was that sample A is perceived as the same as sample B ($H_0: A=B$) if $p > 0,05$. As an alternative hypothesis sample A is perceived different from sample B ($H_a: A \neq B$) if $p \leq 0,05$. Results showed that the three methods did not consistently produce the same range of p values in all the tested product types. On the other hand, all the three methods can be used as a substitute for one another when using milk, cheese, and mixed-flavored yogurts as the observed p values of greater than 0,05 were consistent. Meanwhile, all the remaining products' results were not consistent making it difficult to decide which method is more appropriate than the other.

Figure 30 shows the comparison of two methods to triangle test according to the easiness or difficulty of performing the tests. Duo-trio test was more often described by the participants as easier to perform than triangle test compared to tetrad test. Tetrad test on the other hand was more often described by the participants as more difficult to perform than triangle test compared to duo-trio test. There were several situations where tetrad was described as easier due to the confirmation of choice provided by the fourth sample. In all the sessions at least half of the participants have rated the two new possible methods being as easy or as difficult as the triangle test. This was a good indication that changing or at least alternating the currently used method with other methods was not a bad idea at all. Participants also left some feedback on how they felt about the series of tests. Some found it hard to perform and some found the tests to be interesting and useful (Appendix 4).

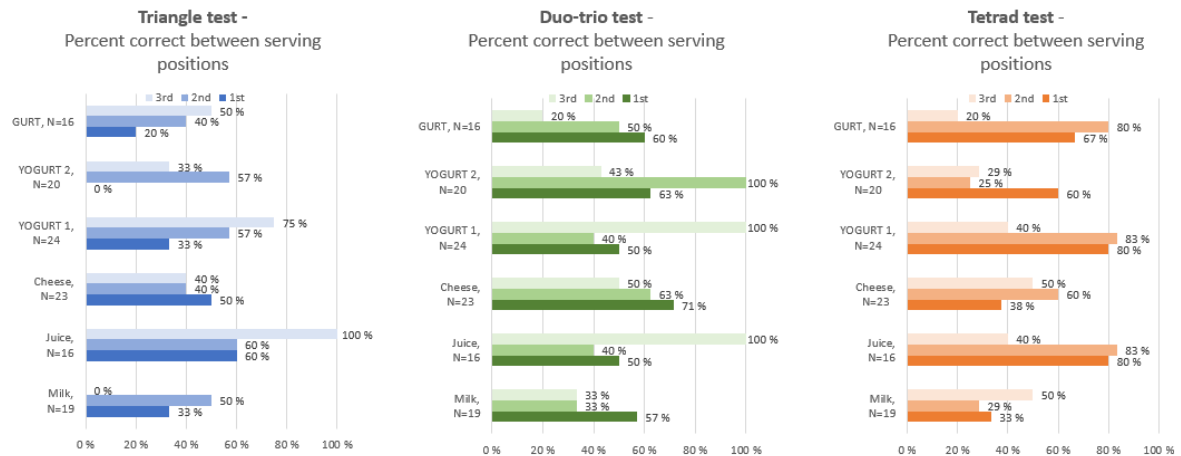
Figure 30. Duo-trio vs. triangle and tetrad vs. triangle



Serving three series of discrimination tests in one session was expected to result in sensory fatigue. Figure 31 shows how serving positions influenced the percent correct answers in each test method. As speculated, the discrimination ability of a tester will decrease by the time the last series was evaluated. The graph in Figure 31 failed to prove the speculation that the percent correct decreases as the serving position of the test series progresses. Not all of the first series tests have the highest percent correct and not all of the third series tests' percent correct answers were the lowest. Results have shown that the one-minute break in between sample series evaluation was useful. Although tetrad test's comments when compared to triangle test were mostly negative due to the addition of a fourth sample,

the percent correct results of tetrad tests showed otherwise instead. The percent correct achieved in tetrad tests were not very low compared to the triangle test. However as a logical result, tetrad served as the last series produced the least percent of correct answers.

Figure 31. Effects of series serving position to percent correct results



Additional analyses were made to provide more information on the results. The illustrations on these results-were presented later in the Appendix section.

Number of total correct answers vs. day of test

Since the tests were offered in several consecutive days in a week, the effect of the test days was also checked. The number of participants in each day varied among the tests, so a direct relation between the day of the week and the number of correct answers produced was not found.

Number of total correct answers vs. time of day

The same comparison was done between the number of correct answers in relation to the time of the day (AM vs. PM). An interesting observation was found, that more correct answers were gathered from the evaluation performed in the daytime (AM) compared to the afternoon (PM).

Number of individual correct answers vs. time duration of the test

Another additional analysis was made to compare the number of correct answers an assessor has made in relation to the time used in performing the test. In one session the maximum number of correct answers a participant can get was three. There was no direct relation between time duration and total correct answers in the case of milk and yogurt samples. This means using more time in the evaluation did not necessarily mean better performance in difference test. Some found the difference in all the three series in yogurt samples at the four-minute evaluation time. The difference in performance level among the participants can be based on either the nature of the samples' sensory difference or the level of expertise and experience in performing the sensory evaluation of the product to be tested. However, when evaluating cheese and juice samples, the more the participant took time to perform the test, the more correct answers were produced. On the contrary, with yogurt samples the longer the time of evaluation, the lower the number of correct answers achieved.

7 Conclusion and recommendations

This study has shown that triangle test is not irreplaceable, it is not always the best option nor the only method of choice for overall difference testing. The method was found to be prone to errors. It is not a one size-fits-all type of method that should always be chosen whenever a difference test implementation is necessary. There are number of other factors to be considered when deciding on the appropriate difference testing method: the number of samples available, the nature and degree of difference between the samples, the timetable of the research, the number of recruited participants, the consistency of the test instructions, the complexity of sample preparation, the difficulty of the discrimination task and the simplicity of communicating the test results.

Attribute-specified discrimination test is a good choice of method to consider whenever the sensory characteristic of concern is identified. Although a change in recipe or processing does not affect a single attribute only, the method is suitable if the objective of the study can be narrowed down specifically to determine the difference only in the attribute of

concern. With a clear objective on proving that the change will make the product less sweet as an example, directional difference testing methods like the n -alternative forced choice (2-AFC or 3-AFC) will help produce better results. As a recommendation by Ennis and Jesionka (1993; 2011), the 2-AFC is the best option for an attribute specified difference testing, as additional number of samples will just create variations in sensory perception affected by fatigue, memory, and adaptation.

For unspecified or overall difference testing, the methods involving more than four samples are considered not practical in many ways. The selected methods constant reference duo-trio and tetrad tests were both good alternatives to triangle test at least according to the product types tested in this experiment. Results showed that triangle, constant reference duo-trio and tetrad tests can be used alternatively when the food type to be evaluated is milk, cheese, or mixed-flavored yogurt. Duo-trio test provided equally significant and reliable results, while only lesser number of samples were required. Tetrad test also provided equally significant and reliable results, while providing the assessors more confidence in their answers with the presence of the fourth sample as a confirmation. Referring to the tables of minimum number of correct responses required to establish significant differences, triangle test requires at least five participants, duo-trio test's minimum participant is six, but the tetrad test's significance can be proven even with only three participants. With the statistical table as a tool, difference test can be conducted using tetrad with fewer participants compared to triangle test.

However, when testing for difference in juice, one-flavored yogurt, and dairy-free yogurt alternative samples the results gained were not consistent. In juice samples, triangle and tetrad methods were more sensitive in determining sensory difference than duo-trio test. In one-flavored yogurt samples, a significant difference was observed only in the triangle test and not in duo-trio test nor tetrad test. In dairy-free yogurt alternative samples, duo-trio and triangle test produced no significant difference in contrast to the tetrad test.

The triangle test's main advantage to the company is familiarity, as it has been used for a long period of time. With many years of experience in organizing triangle tests, this method seemed to be theoretically simple, fast, and easy to implement. Results were also easy to

interpret with the use of the readily available statistical tables. For the assessors, with many years of practice in performing triangle tests, executing the discrimination task is also fast and easy. As assessors get accustomed to the same method over the years, it was observed that performing the discrimination task becomes less efficient. This is because the assessors already know what to expect from the samples, they know exactly what to do, so they tend to perform the triangle test as fast as they can. Asking the assessors to match the samples in duo-trio test or to sort the samples into two groups in tetrad test instead of always asking them to choose the odd sample, stimulated their cognitive decision strategy positively. The assessors gave positive feedbacks that performing the two new methods was interesting, challenging, but fun.

The advantage of constant reference duo-trio test is that the samples needed are less than the triangle test. The reminder of the reference also aids in more stable decision strategy in the assessors' point of view. With a reference sample to match and only two test samples to evaluate, duo-trio is easier to perform by the participants. This was proven by the results of method comparison in easiness or difficulty ratings given by the participants. Initially, duo-trio test was expected to be a lot easier than the triangle, but the results revealed that it was just a little bit easier and mostly as easy or as difficult as the triangle test. The same difficulty rating as the triangle might be affected by the fact that the method was new to the assessors and they are less confident in performing the discrimination task. Nevertheless, organizing duo-trio test is faster and easier while requiring a smaller number of products and involves fewer possible sequence arrangements leading to less sequence variation effect. The only disadvantage of this method is the higher guessing probability ($p=1/2$), which means more participants are needed and a higher number of correct responses is required.

Tetrad test was recently found to be more statistically powerful than triangle and duo-trio tests. It is a new testing method and has been a popular substitute to triangle test. Brewery and alcoholic beverage companies were among the first to replace triangle with tetrad test. Many companies are also considering the switch to tetrad test. The recommended number of participants were often less than a third of those for the triangle test, making this method a better choice for internal difference testing. However, the addition of the fourth sample, while having the same guessing probability of $1/3$ as the triangle test is weighing down the

advantages of this method. It is logical to think that the fourth sample will cause more sensory fatigue when testing strong tasting samples, but the results acquired from this experiment showed otherwise. In method comparison, it was expected that tetrad will be a lot more difficult to perform than the triangle due to the addition of the fourth sample. Although participants' comments on the difficulty of tetrad confirmed the sensory fatigue caused by the fourth sample, the average rating of tetrad's difficulty was the same level as the triangle test. Also, the number of correct answers gained from the tetrad test series was mostly higher compared to triangle test.

The total number of assessors who participated in each test session were insufficient, that concluding a complete replacement of the triangle test does not have a strong basis. There are pros and cons in triangle test, but it is still a valid method for testing sensory difference. Triangle test should not be replaced completely, but other methods will provide alternative options depending on product type and degree of difference. Instead, a table of different discrimination testing methods with their advantages and disadvantages according to the updated ISO and ASTM standards were made for the company for internal reference. Alternatively using these three methods will be a good option to consider in the future, as the assessors start being accustomed to the routinely used triangle test. Using alternative methods will provide training to the internal panel of assessors that will eventually develop their sensory discrimination skills. This study served as a useful learning experience for the employees to practice performing and understanding different discrimination testing methods themselves. This study provided an understanding on recent developments on discrimination tests and a series of practice on performing alternative methods, that will be possible options for the company's difference testing in the future.

A table of recommended sample size (number of participants) required to achieve a certain level of significance and power was made available for use (J. M. Ennis & Jesionka, 2011). By referring to the table, the number of participants needed to assess the samples can be defined depending on the parameters set for data analysis. A deeper understanding on the Thurstonian model for data analysis and its application will be a helpful tool in gaining more confidence in the discrimination testing results. This is an interesting topic to dig deeper as the next step of this study.

8 Acknowledgements

This study will not be possible without the support of my employer Valio Ltd. R&D department in Helsinki. Being a part of the sensory research team made conducting the tests easier as the method comparisons were incorporated with the on-going projects. I was given freedom to creatively implement the experiment using available resources given in these difficult times during the Covid 19 pandemic. I am also grateful to my brilliant M. Sci. colleagues in research and product development Enni Kerola, Hele-Mai Lujanen, Janne Uusi-Rauva, Paula Koivisto and Heli Jokinen, who provided the samples to be tested in this study. With the strict restrictions imposed, I would like to extend my greatest gratitude to all my colleagues in the internal sensory panel who made time to participate in the series of sensory evaluations. Sincere thanks also belong to my colleague M. Sci. Angga Chandrakusuma, who helped me in sample preparation while keeping the sensory laboratory functional. Valuable professional advice from the field's experts Dr. Terhi Pohjanheimo of Aistila and soon-to-be Dr. Maija Greis were also very much appreciated. You both helped me improve the experimental design's structure in the best way possible. To my previous supervisors Dr. Minja Miettinen and Dr. Anu Kaukovirta, thank you for believing in me and for encouraging me to continue my studies. This study will not push forward without the help of my previous adviser Dr. Elina Kytö, who supported me from the start and has given me expert feedback. I was also lucky to receive relevant insights from our team's senior researcher Dr. Sari Puputti, who provided useful improvements on the quality of text in this thesis. To my two advisers Paula Koivisto (M. Sci.) and Tiina Hämäläinen (SVP Product development, fresh dairy, M. Sci.) of Valio Ltd., thank you for accepting the responsibility to take over in guiding this study. This process was also made easier to accomplish by my adviser from the Häme University of Applied Sciences Dr. Tuija Pirttijärvi, your compliments and constructive feedback helped me greatly. To my parents Malou and Bobby Oca and brother Jeffrey John back home, thank you for all the video calls that kept me grounded. I would also like to thank all the difficulties and challenges along the way, that helped me go beyond my limits. Without these, I would not have the chance to improve myself. Above all, none of this will be made into reality without the unconditional love of my husband Robert and children Maria Mnorje and Marcus Davion, who were with me during the ups and downs of this whole process. From you all, I gained confidence that I can do this. Thank you.

References

- Abdou, B., Papa, M. D. D. S., Adama, D., Laure, T., Mame, S. M., Ndeye, S. C., Nicole, I. D., Salimata, W., & Philippe, D. (2018). Sensory evaluation and consumer acceptability of orange-fleshed sweet potato by lactating women and their children (. *African Journal of Food Science*, 12(11). <https://doi.org/10.5897/AJFS2018.1730>
- Adawiyah, D. R., Guntari, L., Smaratika, V. S., & Lince. (2020). A comparison of tetrad and triangle test: Case study on sweetener products using consumer panels. *IOP Conference Series: Earth and Environmental Science*, 443(1). <https://doi.org/10.1088/1755-1315/443/1/012090>
- Adjei, M. Y. B. (2017). Applications and Limitations of Discrimination Testing. In *Discrimination Testing in Sensory Science: A Practical Handbook* (pp. 85–105). Elsevier. <https://doi.org/10.1016/B978-0-08-101009-9.00004-6>
- Amerine, M. A., Pangborn, R. M., & Roessler, E. B. (1965). *Principles of Sensory Evaluation of Food*. Elsevier. <https://doi.org/10.1016/C2013-0-08103-0>
- ASTM E3009-15e1. (2015). ASTM E3009-15e1, Standard Test Method for Sensory Analysis—Tetrad Test. In *SFS Online*. SFS Online. <https://doi.org/10.1520/E3009-15E01>
- Basker, D. (1980). Polygonal and polyhedral taste testing. *Journal of Food Quality*, 3(1). <https://doi.org/10.1111/j.1745-4557.1980.tb00682.x>
- Bi, J. (2015). *Sensory Discrimination Tests and Measurements*. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118994863>
- Bi, J. (2020). A new form of the psychometric function for the unspecified tetrad. *Food Quality and Preference*, 82.
- Bissmeyer, D. (2019). Validating the Sensitivity of the Beer Tetrad Test as Compared with the Beer Triangle Test (A Follow-Up Study): An American Society of Brewing Chemists

Technical Committee Report. *Journal of the American Society of Brewing Chemists*, 77(3). <https://doi.org/10.1080/03610470.2019.1619323>

Brockhoff, P. B., & Christensen, R. H. B. (2010). Thurstonian models for sensory discrimination tests as generalized linear models. *Food Quality and Preference*, 21(3), 330–338.

Carr, T. (2021). *Talk: A Trip Through Sensory Science -- From Methods to Measures to Models*. <http://www.pangbornsymposium.com/bio-carr.asp>

Chaves, K. F., Wahanik, A. L., Paludo, M. C., Toledo, B. I., Leme, A. M. V., Orelli Junior, A. A., & Behrens, J. H. (2020). Tetrad vs. triangle test: A case study with Brazilian guarana soft drink. *Research, Society and Development*, 9(4). <https://doi.org/10.33448/rsd-v9i4.3049>

Crofton, E. C., Botinestean, C., Fenelon, M., & Gallagher, E. (2019). Potential applications for virtual and augmented reality technologies in sensory science. *Innovative Food Science & Emerging Technologies*, 56, 102178. <https://doi.org/10.1016/j.IFSET.2019.102178>

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies*, 8(4). <https://doi.org/10.1111/j.1745-459X.1993.tb00225.x>

Ennis, J. M. (2012). Guiding the switch from triangle testing to tetrad testing. *Journal of Sensory Studies*, 27(4), 223–231. <https://doi.org/10.1111/j.1745-459X.2012.00386.x>

Ennis, J. M., Ennis, D. M., Yip, D., & O'Mahony, M. (1998). Thurstonian models for variants of the method of tetrads. *British Journal of Mathematical and Statistical Psychology*, 51(2), 205–215. <https://doi.org/10.1111/J.2044-8317.1998.TB00677.X>

Ennis, J. M., & Jesionka, V. (2011). The Power Of Sensory Discrimination Methods Revisited. *Journal of Sensory Studies*, 26(5), 371–382. <https://doi.org/10.1111/j.1745-459X.2011.00353.x>

- Frijters, J. E. R. (1984). Sensory difference testing and the measurement of sensory discriminability. In J.R. Piggott (Ed.), *Sensory analysis of foods* (pp. 117–140).
- Fuentes, S., Tongson, E., & Gonzalez Viejo, C. (2021). Novel digital technologies implemented in sensory science and consumer perception. *Current Opinion in Food Science*, *41*, 99–106. <https://doi.org/10.1016/J.COFS.2021.03.014>
- Garcia, K., Ennis, J. M., & Prinyawiwatkul, W. (2012a). A large-scale experimental comparison of the tetrad and triangle tests in children. *Journal of Sensory Studies*, *27*(4). <https://doi.org/10.1111/j.1745-459X.2012.00385.x>
- Garcia, K., Ennis, J. M., & Prinyawiwatkul, W. (2012b). A large-scale experimental comparison of the tetrad and triangle tests in children. *Journal of Sensory Studies*, *27*(4), 217–222. <https://doi.org/10.1111/j.1745-459X.2012.00385.x>
- Gonzalez Viejo, C., Torrico, D. D., Dunshea, F. R., & Fuentes, S. (2019). Emerging Technologies Based on Artificial Intelligence to Assess the Quality and Consumer Preference of Beverages. *Beverages*, *5*(4), 62. <https://doi.org/10.3390/BEVERAGES5040062>
- In-Ah Kim, Ji-Young Yoon, & Hye-Seong Lee. (2015). Measurement of consumers' sensory discrimination and preference: Efficiency of preference-difference test utilizing the 3-point preference test precedes the same-different test. *Food Sci Biotechnology*, *24*, 1355–1362. <https://doi.org/10.1007/s10068-015-0174-0>
- Ishii, R., O'Mahony, M., & Rousseau, B. (2014). Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. *Food Quality and Preference*, *31*(1), 49–55. <https://doi.org/10.1016/J.FOODQUAL.2013.07.007>
- ISO 4120:2021. (2021). ISO 4120:2021(en) Sensory analysis — Methodology — Triangle test. In *SFS Online*. International Organization for Standardization. <https://sales.sfs.fi/fi/index/tuotteet/SFS/CENISO/ID2/4/979431.html.stx>

ISO 6658:2017. (2017). ISO 6658:2017(en) Sensory analysis — Methodology — General guidance. In *SFS Online*. SFS Online.

<https://sales.sfs.fi/fi/index/tuotteet/ISO/ISO/ID9998/6/525726.html.stx>

ISO 10399:2017. (2017). ISO 10399:2017(en) Sensory analysis — Methodology — Duo-trio test. In *SFS Online*. SFS Online.

<https://sales.sfs.fi/fi/index/tuotteet/ISO/ISO/ID9998/1/628849.html.stx>

Jeong, Y. N., Kang, B. A., Jeong, M. J., Song, M. J., Hautus, M. J., & Lee, H. S. (2016). Sensory discrimination by consumers of multiple stimuli from a reference: Stimulus configuration in A-Not AR and constant-ref. duo-trio superior to triangle and unspecified tetrad? *Food Quality and Preference*, *47*, 10–22.

Kim, M. A., Chae, J. E., van Hout, D., & Lee, H. S. (2014a). Higher performance of constant-reference duo-trio test incorporating affective reference framing in comparison with triangle test. *Food Quality and Preference*, *32*.

<https://doi.org/10.1016/j.foodqual.2013.08.013>

Kim, M. A., Chae, J. E., van Hout, D., & Lee, H. S. (2014b). Higher performance of constant-reference duo-trio test incorporating affective reference framing in comparison with triangle test. *Food Quality and Preference*, *32*, 113–125.

Kim, M. A., Sim, H. M., & Lee, H. S. (2015). Affective discrimination methodology: Determination and use of a consumer-relevant sensory difference for food quality maintenance. *Food Research International*, *70*, 47–54.

<https://doi.org/10.1016/J.FOODRES.2015.01.027>

Kuesten, C. L. (2001). Sequential use of the triangle, 2-AC, 2-AFC, and same-different methods applied to a cost-reduction effort: consumer learning acquired throughout testing and influence on preference judgements. *Food Quality and Preference*, *12*(5–7), 447–455. [https://doi.org/10.1016/S0950-3293\(01\)00036-2](https://doi.org/10.1016/S0950-3293(01)00036-2)

- Lawless, H. T., & Heymann, H. (2010). *Sensory Evaluation of Food* (2nd ed.). Springer New York. <https://doi.org/10.1007/978-1-4419-6488-5>
- Linander, C. B., Christensen, R. H. B., Cleaver, G., & Brockhoff, P. B. (2019). Individual differences in replicated multi-product experiments with Thurstonian mixed models for binary paired comparison data. *Food Quality and Preference*, *75*, 220–229.
- Maciel, G. M., Poulsen, N. A., Larsen, M. K., Kidmose, U., Gaillard, C., Sehested, J., & Larsen, L. B. (2016). Good sensory quality and cheesemaking properties in milk from Holstein cows managed for an 18-month calving interval. *Journal of Dairy Science*, *99*(11). <https://doi.org/10.3168/jds.2016-10958>
- Martens, M. (1999). A philosophy for sensory science. *Food Quality and Preference*, *10*(4–5), 233–244. [https://doi.org/10.1016/S0950-3293\(99\)00024-5](https://doi.org/10.1016/S0950-3293(99)00024-5)
- McClure, S., & Lawless, H. T. (2010). Comparison of the triangle and a self-defined two alternative forced choice test. *Food Quality and Preference*, *21*(5), 547–552. <https://doi.org/10.1016/J.FOODQUAL.2010.02.005>
- Meilgaard, M. C., Ceville, G. V., & Carr, B. T. (2006). Overall Difference Tests: Does a Sensory Difference Exist between Samples? In *Sensory Evaluation Techniques* (4th ed., pp. 63–104). CRC Press. <https://doi.org/10.1201/b16452>
- Motoki, K., Saito, T., & Onuma, T. (2021). Eye-tracking research on sensory and consumer science: A review, pitfalls and future directions. *Food Research International*, *145*, 110389. <https://doi.org/10.1016/J.FOODRES.2021.110389>
- O'Mahony, M. (1995a). Who told you the triangle test was simple? *Food Quality and Preference*, *6*(4), 227–238. [https://doi.org/10.1016/0950-3293\(95\)00022-4](https://doi.org/10.1016/0950-3293(95)00022-4)
- O'Mahony, M. (1995b). Who told you the triangle was simple? In *Food Quality and Preference* (Vol. 6).

- O'Mahony, M., & Rousseau, B. (2003). Discrimination testing: a few ideas, old and new. *Food Quality and Preference*, *14*(2), 157–164. [https://doi.org/10.1016/S0950-3293\(02\)00109-X](https://doi.org/10.1016/S0950-3293(02)00109-X)
- RedJade Sensory Solutions. (n.d.). *RedJade Sensory Software*. Retrieved September 9, 2021, from <https://redjade.net/>
- Roessler, E. B., Warren, J., & Guymon, J. F. (2006). Significance in triangular taste tests. *Journal of Food Science*, *13*(6), 503–505. <https://doi.org/10.1111/j.1365-2621.1948.tb16650.x>
- Ross, C. F. (2021). Considerations of the use of the electronic tongue in sensory science. *Current Opinion in Food Science*, *40*, 87–93. <https://doi.org/10.1016/J.COFS.2021.01.011>
- Rousseau, B., & O'Mahony, M. (2000). Investigation of the effect of within-trial retasting and comparison of the dual-pair, same-different and triangle paradigms. *Food Quality and Preference*, *11*(6), 457–464.
- Sanderson, T. (2017). Tetrad Test. In *Discrimination Testing in Sensory Science: A Practical Handbook* (pp. 183–195). Elsevier. <https://doi.org/10.1016/B978-0-08-101009-9.00009-5>
- SIMS. (n.d.). *SIMS Sensory software as a cloud service*. Retrieved December 11, 2020, from <https://www.sims2000.com/sample8.asp>
- Theses, M., & Carlisle, S. L. (2014). *Trace: Tennessee Research and Creative Exchange Comparison of Triangle and Tetrad Discrimination Methodology in Applied, Industrial Manner*. https://trace.tennessee.edu/utk_gradthes/2798
- Thurstone, L. L. (1927). Psychophysical analysis. *38*(3), 368–389. *The American Journal of Psychology*, *38*(3), 368–389.

- Tran, C. T. H., Nguyen, P. Q., Pham, Q. T., & Nguyen, D. H. (2014). Is triangle test more powerful than tetrad test in case of high alcoholic beverages? *Science and Technology Development Journal*, 17(3). <https://doi.org/10.32508/stdj.v17i3.1499>
- van Hout, D. H. A. (2014). *Measuring Meaningful Differences*. <https://repub.eur.nl/pub/50387/>
- Worch, T., & Delcher, R. (2013a). A Practical Guideline for Discrimination Testing Combining both the Proportion of Discriminators and Thurstonian Approaches. *Journal of Sensory Studies*, 28(5). <https://doi.org/10.1111/joss.12065>
- Worch, T., & Delcher, R. (2013b). A Practical Guideline for Discrimination Testing Combining both the Proportion of Discriminators and Thurstonian Approaches. *Journal of Sensory Studies*, 28(5), 396–404. <https://doi.org/10.1111/joss.12065>
- Xia, Y., Zhang, J., Zhang, X., Ishii, R., Zhong, F., & O'Mahony, M. (2015). Tetrads, triads and pairs: Experiments in self-specification. *Food Quality and Preference*, 40(PA), 97–105. <https://doi.org/10.1016/J.FOODQUAL.2014.09.005>

Appendix 1: Justifications on method selection for comparison

Selection of duo-trio testing method

Reason 1 based on (van Hout, 2014, p. 20)

WHY TEST MATCHING TECHNIQUE WITH CONSTANT REFERENCE DUO-TRIO? →1/4

Table 2.2 Task and response structures of seven popular test methods

Test method	Instructions	Reference product	Reminder of the reference product	Number of samples in a trial	Number of responses in a trial	Sureness rating	Literature references describing the methods
A-Not A	Is this product A (the reference) or Not A?	A	No	1	1	recommended	O'Mahony 1979a, 1982; Lee, Van Hout and O'Mahony 2006; Lee and Van Hout 2009
A-Not AR	Is this product A (the reference) or Not A?	A	Yes	2	1	recommended	O'Mahony 1979, 1982; Lee and Van Hout 2009
Same-different	Are these two products the same or different?	no reference	No	2	1	recommended	Rousseau, Meyer, and O'Mahony, 1998; Lee, Van Hout, and Hautus, 2007b
2-AFC	Which of the two products is A (the reference)?	A	No	2	1	optional	Meilgaard, Civille, and Carr, 1999;
★ 2-AFCR (Duo-trio fixed reference)	Which of the two products is A (the reference)?	A	Yes	3	1	optional	Hautus, Shepherd and Peng 2011;
Duo-trio (balanced reference)	Which of the two products is the reference (A or B)?	A or B	Yes	3	1	optional	Lee and Kim, 2008
★ Triangle	Which of the three products is different from the other two?	no reference	No	3	1	optional	Meilgaard, Civille, and Carr, 1999; Lee and Kim 2008

DANIELLE VAN HOUT
Measuring Meaningful Differences
 Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling

Reason number 1

MATCHING:
Constant reference duo-trio test

- Same sample number as the triangle test

Reason 2 based on (van Hout, 2014, p. 22)

WHY TEST MATCHING TECHNIQUE WITH CONSTANT REFERENCE DUO-TRIO? →2/4

Table 2.3 Sequences and estimated strengths of the sequence effects in the seven difference test methods and corresponding literature references.

Test method	Number of possible sequences	Types of sequences Reference product between brackets (A)	Variance caused by sequence effects (1=lowest)	Literature references investigating sequence effects
A-Not A	2	A, B	1	-
A-Not AR	2	(A)A, (A)B	2	-
Same-different	4	AA, AB, BA, BB	3	Rousseau, Meyer, and O'Mahony, 1997; Santosa and O'Mahony, 2008;
2-AFC	2	AB, BA	2	Santosa, Hautus, O'Mahony, 2011 Dessirier and O'Mahony, 1999;
★ 2-AFCR	2	(A)AB, (A)BA	4	Lee and Kim, 2008; Kim, et. Al., 2013;
Duo-trio	4	(A)AB, (A)BA, (B)AB, (B)BA	5	Lee and Kim, 2008; Kim, et. Al., 2013; Rousseau, Meyer, and O'Mahony, 1997
★ Triangle	6	AAB, ABA, ABB, BAA, BAB, BBA	6	Kim, et. Al., 2013; Rousseau, Meyer, and O'Mahony, 1997; O'Mahony 1995b

DANIELLE VAN HOUT
Measuring Meaningful Differences
 Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling

Reason number 2

MATCHING:
Constant reference duo-trio test

- Fewer possible sequences: faster and easier to organize, less chance of sequence errors

Reason 3 based on (van Hout, 2014, p. 23)

WHY TEST MATCHING TECHNIQUE WITH CONSTANT REFERENCE DUO-TRIO? →3/4

Table 2.4 Most likely decision strategies of seven difference test methods and corresponding literature references.

Test method	Cognitive decision strategies	Effects of cognitive strategy changes	Literature references investigating cognitive decision strategies in sensory tests
A-Not A	beta	When subjects are sufficiently familiar with the reference product they will use the beta strategy	Hautus et al. 2009; O' Mahony and Hautus 2008
A-Not AR	Usually beta, sometimes tau - comparison of distances	Strategy depends on how a subject uses the information from the reminder, differences between subjects introduces variance	Hautus et al. 2009, 2011a, Stocks et al 2013, 2014.
Same-different	Mostly tau - comparison of distances	Strategy changes to beta possible when introduction consists of familiarization with beta task	Santosa and O' Mahony, 2008; Santosa et al, 2011; Rousseau, Stroh, and O' Mahony, 2002 ; Lee et al. 2007a, 2007b; Rousseau and O'Mahony 2000 Hautus et al., 2009, 2011a, 2011b
2-AFC	beta	When subjects are familiar with the reference they either use beta or tau-optimal, and these two strategies lead to the same results	
★ 2-AFCR	mostly beta, or tau-optimal (skimming), sometimes tau comparison of distances	When subjects are familiar with the reference they either use beta or tau-optimal, and these two strategies lead to the same results	Hautus et al., 2009, 2011a, 2011b; Kim, Lee, and Lee, 2010; Kim, Chae, Van Hout, and Lee, 2014
Duo-trio	tau - comparison of distances	Re-tasting can cause strategy shifts to more optimal strategies, difficult to model free retesting with SDT.	Kim, Lee and Lee, 2010; Lee and Kim 2008; Kim and Lee, 2012
★ Triangle	tau, or 'triangle beta'	Re-tasting can cause some strategy shifts to triangle beta. This makes it complicated to model with SDT and therefore results in inaccurate d' estimates	Rousseau 2001; O'Mahony 1995b; Rousseau and O'Mahony 2000

Reason number 3

DANIELLE VAN HOUT
Measuring Meaningful Differences

Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling

MATCHING:

Constant reference duo-trio test

- The reference serving only as a reminder lessens the number of samples to be compared, thus inhibiting changes in decision strategies

Although the guessing probability is only 1/3, which is less than the triangle test's 1/3, the method is concluded to have more testing power (reliability of results) than the triangle test.

2 1

Reason 4 based on (van Hout, 2014, p. 25)

WHY TEST MATCHING TECHNIQUE WITH CONSTANT REFERENCE DUO-TRIO? →4/4

Table 2.5 Optimal performance levels of the seven difference test methods, and number of trials required to detect various sizes of sensory differences (d'), with a power of 0.8 and alpha = 0.05, and corresponding literature references.

Test method	Optimal performance level	Number of trials required for detecting differences , with alpha=0.05 and power= 0.8				Literature references, comparing test power and sensitivity
		d' =0.5	d' =1	d' =1.5	d' =2	
A-Not A	d'=1-2	91	29	17	10	Bi and Ennis, 2001a, 2001b
A-Not AR	d'=1-2	91	29	17	10	Bi and Ennis, 2001a, 2001b; Hautus et al., 2009
Same-different	d'=1.5-2.5	2825	220	57	23	Rousseau, Meyer, and O'Mahony, 1998
2-AFC	d'=0.5-1.5	89	26	13	8	Ennis, 1993; Ennis and Jesionka, 2011
★ 2-AFCR	d'=0.5-1.5	89	26	13	8	Hautus, Van Hout, and Lee, 2009; Hautus, Shepperd, and Peng, 2011a
Duo-trio	d'=2-3	3160	241	65	28	Ennis, 1993; Ennis and Jesionka 2011
★ Triangle	d'=2-4	2825	220	57	23	Ennis, 1993; Ennis and Jesionka 2011

Reason number 4

DANIELLE VAN HOUT
Measuring Meaningful Differences

Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling

MATCHING:

Constant reference duo-trio test

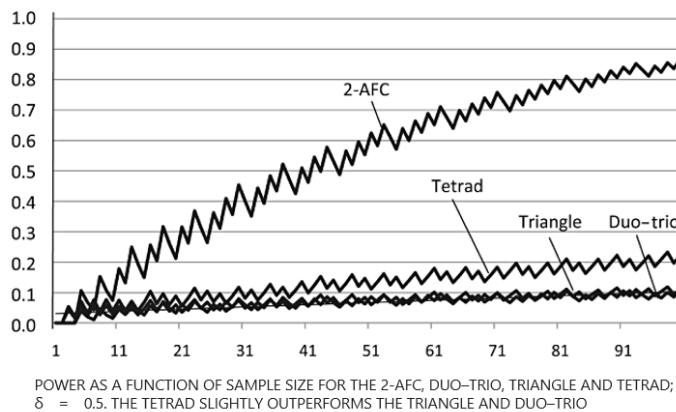
- More sensitive in detecting differences as seen with fewer required number of trials compared to triangle test

2 2

Selection of tetrad testing method

Reason 5 based on (J. M. Ennis & Jesionka, 2011, p. 378)

WHY TEST SORTING TECHNIQUE WITH THE TETRAD? → 1/2



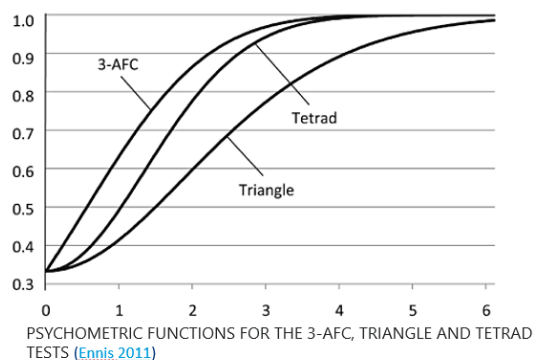
Reason number 5

"We found that the (unspecified) tetrad test was remarkably powerful in comparison to the duo-trio and triangle tests. This fact was reinforced in the sample size tables that we updated; in these tables, the recommended sample sizes for the tetrad test were often less than a third of those for the triangle test". (Ennis 2011)



Reason 6 based on (J. M. Ennis & Jesionka, 2011, p. 377)

WHY TEST SORTING TECHNIQUE WITH THE TETRAD? → 2/2



Reason number 6

"Since the guessing probability of the tetrad test is $1/3$, it is very interesting to compare this function to the corresponding psychometric functions for triangle test.

It is remarkable to note that the tetrad test, even though it does not specify an attribute, has a psychometric function that is intermediate between those for the triangle and the 3-AFC. From this result, we predict higher power for the tetrad test than for the triangle. This higher power will translate into lower sample size requirements. (Ennis 2011)



Appendix 2: Test questionnaire



Survey Created Using RedJade Software

Questionnaire | Page 1

Taustatieto erotustestistä

Mikä on erotustesti?

- Erotustestien avulla selvitetään, vaikuttaako muutos (mm. raaka-aine, pakkaus, valmistusprosessi, säilyvyyden olosuhde) tuotteeseen ja eroavatko näytteet toisistaan aistinvaraisesti.
- Erotustestissä ei ole väärä tai oikea vastausta.
- Erotustestit käytetään myös yhdisteiden kynnysarvojen määrittämisessä, asiantuntijaraadin koulutuksessa tai menetelmän testaamisessa.

Mitkä ovat erilaisia erotustestejä?

- Kolmitesti
- Pari-kolmitesti
- Neliötesti
- Parivertailutesti
- Kaksi viidestä -testi
- On-ei ole -testi
- Suunnattu parivertailutesti, jossa aistinvarainen ominaisuus on määritetty
- Suunnattu kolmitesti, kuten suunnattu parivertailussa, mutta kolmen vaihtoehdon pakkovalintatesti

Taustatieto tämän päivän erotustesteistä

Kolmitesti

- Yleisin käytetty erotustesti
- Erotustestimenetelmä käytössä Valiolla
- Arvioijalle esitetään **kolme näytettä**, joista kaksi on samanlaista ja yksi poikkeava
- Arvioijan tehtävänä on tunnistaa poikkeava näyte sarjasta
- Arvausmahdollisuus 1/3

Pari-kolmitesti

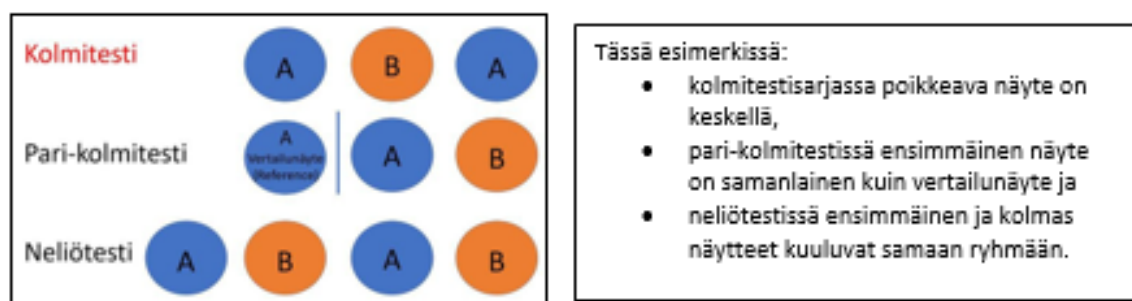
- Arvioijalle esitetään ensin **vertailunäyte** ja sen jälkeen **kaksi näytettä**, joista toinen on samanlainen kuin vertailunäyte
- Arvioijan tehtävänä on tunnistaa vertailun kaltainen näyte
- Arvausmahdollisuus 1/2

Neliötesti

- Arvioijalle esitetään samanaikaisesti **neljä näytettä**, joista kaksi on keskenään samanlaisia.
- Arvioijan tehtävänä on jakaa näytteet kahteen ryhmään
- Arvausmahdollisuus 1/3

Jokaiselle arvioijalle esitetään kolme näytesarjaa, joista kolmitesti, pari-kolmitesti ja neliötesti ovat satunnaisessa järjestyksessä. Näytesarjojen välissä on 1 minuutin tauko. Huuhto suusi hyvin vedellä sarjojen välissä ja/tai syö vesikeksejä makuaistin palauttamiseksi. Testin lopussa näytetään näytekoodit, joilla voit tarkistaa löysitkö eroja näytteiden välillä. Kirjoita omat vastaukset paperille tarkistusta varten.

Esimerkkikuva tarjottimen näytejärjestyksestä alla (Näyte A vs. näyte B):





Survey Created Using RedJade Software

Questionnaire Page 1

1 Kolmitesti (I rangle test)

Huuho suusi vedellä ennen arviointia.

Arvioitavanasi on kolme näytettä, joista kaksi on samanlaista ja yksi on poikkeava.

Tehtävänäsi on maistaa jokaista näytettä lomakkeen mukaisessa järjestyksessä vasemmalta oikealle ja valita poikkeava näyte.

Ero voi olla ulkonäössä, rakenteessa, suutuntumassa ja/tai maussa.

Arvaa, mikäli et ole varma vastauksestasi. Huuhdo suusi vedellä näytteiden välissä. Näytteitä saa maistaa tarvittaessa uudelleen, mutta vain samassa järjestyksessä kuin lomakkeessa.

näyte ###



näyte ###



näyte ###



2 Ilmoita eron suuruus seuraavan asteikon mukaisesti (0-3)

0 = ei eroa



1 = pieni, juuri havaittava ero



2 = selvä ero



3 = erittäin selvä ero



NOTE: Only answer this question if on question #2 of questionnaire page 1 your answer was one of the following: "1 = pieni, juuri havaittava ero" "2 = selvä ero" "3 = erittäin selvä ero"

3 Miten poikkeava näyte eroaa kahdesta samanlaisesta näytteestä? Kuvaile eroa.



Survey Created Using RedJade Software

Questionnaire Page 1

1 Pari-kolmitesti (Duo-trio test)

Huuhto suusi vedellä ennen arviointia.

Arvioitavanasia on kolme näytettä, joista ensimmäinen on vertailunäyte (REFERENCE) ja kaksi testinäytettä. Maista ensin vertailunäytettä.

Tehtävänäsi on maistaa testinäytteitä lomakkeen mukaisessa järjestyksessä vasemmalta oikealle ja valita vertailunäytteen kaltainen näyte.

Ero voi olla ulkonäössä, rakenteessa, suutuntumassa ja/tai maussa.

Arvaa, mikäli et ole varma vastauksestasi. Huuhto suusi vedellä näytteiden välissä. Näytteitä saa maistaa tarvittaessa uudestaan, mutta vain samassa järjestyksessä kuin lomakkeessa.

REFERENCE

näyte ###

näyte ###



2 Ilmoita kahden testinäytteiden välisen eron suuruus seuraavan asteikon mukaisesti (0-3).

- | | | | |
|-----------------------|-------------------------------|-----------------------|-----------------------|
| 0=ei eroa/arvasin | 1=pieni, juuri havaittava ero | 2=selvä ero | 3=erittäin selvä ero |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

NOTE: Only answer this question if on question #2 of questionnaire page 1 your answer was one of the following: "1=pieni, juuri havaittava ero" "2=selvä ero" "3=erittäin selvä ero"

3 Eron kuvaus



Survey Created Using RedJade Software

Questionnaire Page 1

Please select between 2 and 2 responses

1 Neliötesti (Tetrad test)

Huuhto suusi vedellä ennen arviointia.

Arvioitavanasi on neljä näytettä, joista kaksi on keskenään samanlaisia.

Tehtävänäsi on maistaa jokaista näytettä lomakkeen mukaisessa järjestyksessä vasemmalta oikealle ja jakaa näytteet kahteen ryhmään (Ryhmä A ja Ryhmä B), joissa on kaksi keskenään samanlaista näytettä. Lajittele näytteet vetämällä näytekoodit kahteen ryhmään A tai B.

Ero voi olla ulkonäössä, rakenteessa, suutuntumassa ja/tai maussa.

Arvaa, mikäli et ole varma vastauksestasi. Huuhto suusi vedellä näytteiden välissä. Näytteitä saa maistaa tarvittaessa uudestaan, mutta vain samassa järjestyksessä kuin lomakkeessa.

näyte ###	näyte ###	näyte ###	näyte ###
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ryhmä A		Ryhmä B	
<input type="text"/>		<input type="text"/>	

2 Ilmoita kahden ryhmän välisen eron suuruus seuraavan asteikon mukaisesti (0-3)

0= ei eroa	1= pieni, juuri havaittava ero	2= selvä ero	3= erittäin selvä ero
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

NOTE: Only answer this question if on question #2 of questionnaire page 1 your answer was one of the following: "1= pieni, juuri havaittava ero" "2= selvä ero" "3= erittäin selvä ero"

2 Eron kuvaus:



Survey Created Using RedJade Software

Questionnaire Page 1

1 **Pari-kolmitesti vs. kolmitesti**

Mitä mieltä olet **pari-kolmitestin** suorittamisesta verrattuna kolmitestiin?

paljon helpompaa kuin kolmitesti	hieman helpompaa kuin kolmitesti	yhtä helppo tai yhtä vaikea kuin kolmitesti	hieman vaikeampaa kuin kolmitesti	paljon vaikeampaa kuin kolmitesti
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2 **Neliötesti vs. kolmitesti**

Mitä mieltä olet **neliötestin** suorittamisesta verrattuna kolmitestiin?

paljon helpompaa kuin kolmitesti	hieman helpompaa kuin kolmitesti	yhtä helppo tai yhtä vaikea kuin kolmitesti	hieman vaikeampaa kuin kolmitesti	paljon vaikeampaa kuin kolmitesti
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

NOTE: Only answer this question if on question #1 of questionnaire page 1 your answer was one of the following: "paljon helpompaa kuin kolmitesti" "hieman helpompaa kuin kolmitesti"

3 Perustelisitko vastauksesi, miksi **pari-kolmitesti** on **helpompi** kuin kolmitesti?

NOTE: Only answer this question if on question #1 of questionnaire page 1 your answer was one of the following: "hieman vaikeampaa kuin kolmitesti" "paljon vaikeampaa kuin kolmitesti"

3 Perustelisitko vastauksesi, miksi **pari-kolmitesti** on **vaikeampi** kuin kolmitesti?

NOTE: Only answer this question if on question #2 of questionnaire page 1 your answer was one of the following: "paljon helpompaa kuin kolmitesti" "hieman helpompaa kuin kolmitesti"

4 Perustelisitko vastauksesi, miksi **neliötesti** on **helpompi** kuin kolmitesti?

NOTE: Only answer this question if on question #2 of questionnaire page 1 your answer was one of the following: "hieman vaikeampaa kuin kolmitesti" "paljon vaikeampaa kuin kolmitesti"

5 Perustelisitko vastauksesi, miksi **neliötesti** on **vaikeampi** kuin kolmitesti?



Survey Created Using RedJade Software

Questionnaire

Taustatietosi

1 Sukupuolesi

- Nainen
- Mies
- Muu

2 Ikäryhmäsi

- Alle 25 vuotta
- 25-34 vuotta
- 35-44 vuotta
- 45-54 vuotta
- 55-64 vuotta
- Yli 65 vuotta

3 Lopuksi voit vielä halutessasi jättää muita kommentteja, ajatuksia, kehitysehdotuksia ja/tai ideoita.

Appendix 3. Additional results

A. Degree of difference

MILK

Eron suuruus eron havainneiden keskuudessa, vain kolmitestissä oikein vastanneet:

	KOLMITESTI VRT vs. KOE 2 Frekvenssi N=5	PARI- KOLMITESTI VRT vs. KOE 2 Frekvenssi N=8	NELIÖTESTI VRT vs. KOE 2 Frekvenssi N=7
0 = ei eroa	2	2	3
1 = pieni, juuri havaittava ero	3	6	4
2 = selvä ero	0	0	0
3 = erittäin selvä ero	0	0	0
keskiarvo	0,6	0,8	0,6

Eron kuvailu, (jos eron suuruus on > 0):

	VRT	KOE 2	Näytteiden eron kuvailu
Kolmitesti, N=3	kitkerämpi/ karvaampi	vähemmän makea, puhtain maku	
Pari- kolmitesti, N=6	sivumaku, ei puhdas	hieman kitkerämpi, laimeamman makuinen	N/A (2), makeus
Neliötesti, N=4	tuoreempi maku	voimakkaamman kitkerää/karvasta, laimeampia	makeus

JUICE

Eron suuruus eron havainneiden keskuudessa, vain kolmitestissä oikein vastanneet:

	KOLMITESTI VRT vs. KOE Frekvenssi N=12	NELIÖTESTI VRT vs. KOE Frekvenssi N=11	PARI-KOLMITESTI VRT vs. KOE Frekvenssi N=10
0 = ei eroa	2	2	4
1 = pieni, juuri havaittava ero	9	7	5
2 = selvä ero	1	2	1
3 = erittäin selvä ero	0	0	0
keskiarvo	0,9	1,0	0,7

Eron kuvailu, jos eron suuruus on > 0:

Erotus- testi	Näytteiden eron kuvailu
Kolmitesti, N=10	<i>poikkeavan näytteen ero:</i> VRT (n=5) puhtaampi makeampi maku, <i>happoisempi/hapokkaampi (2), laimeampi, n/a</i> KOE (n=5) imelämpi maku, mutten ole tästä varma. On kyllä tuskin havaittavia eroavaisuuksia, mieta, <i>makeampi (2), happamuus</i>
Neliötesti, N=9	<i>kahden ryhmän ero:</i> <i>happamuus (2), makeus (6), maun voimakkuus (1)</i>
Pari- kolmitesti, N=6	<i>kahden testinäytteen ero:</i> <i>happamuus (4), maun voimakkuus, makeus (2)</i>

CHEESE

KOLMITESTI - ERON SUURUUS (VAIN KOLMITESTISSÄ OIKEIN VASTANNEET, N=10)

Eron suuruus vain eron havainneiden keskuudessa:

	VRT vs. KOE Frekvenssi N=10
0 = ei eroa	0
1 = pieni, juuri havaittava ero	5
2 = selvä ero	4
3 = erittäin selvä ero	1
keskiarvo	1,6

poikkeava näyte sarjassa	Eron kuvailu N=10, (jos eron suuruus on > 0)
VRT N=2	aavistuksen makeahko, voimakkaampi
KOE N=8	<i>Rakenne on paljon pehmeämpi (4), ei niin kiinteä kuin muut, muut näytteet ovat kovia, murenevampi.</i> <i>Voimakkaampi maku, pehmeämpi jälkimaku, ei niin karvas, kitkerämpi maku, enemmän voihapon maku, maukkaampi, muut näytteet mauttomampia.</i> <i>Kellertävämpi väri.</i>

YOGURT1

**ERON SUURUUS –
VAIN KOLMITESTISSÄ OIKEIN VASTANNEET, N=13**Eron suuruus vain eron havainneiden keskuudessa:

	VRT vs. KOE Frekvenssi N=13
0 = ei eroa	5
1 = pieni, juuri havaittava ero	8
2 = selvä ero	0
3 = erittäin selvä ero	0
keskiarvo	0,6



Poikkeava näyte sarjassa	Eron kuvailu N=8, (jos eron suuruus on > 0)
VRT N=4	Vähemmän heroittunut ja hieman makeampi, Paksumpi rakenne, Hieman kirpeämpi, Hieman löysempi rakenne, ihan vähän vähemmän hapan
KOE N=4	Pistävä jälkimaku, Hieman happamampi ja voimakkaampi tuoksu, Hieman erilainen maultaan, Hieman makeampi

YOGURT 2

**ERON SUURUUS –
VAIN KOLMITESTISSÄ OIKEIN VASTANNEET, N=6**Eron suuruus vain eron havainneiden keskuudessa:

	VRT vs. KOE Frekvenssi N=6
0 = ei eroa	3
1 = pieni, juuri havaittava ero	3
2 = selvä ero	0
3 = erittäin selvä ero	0
keskiarvo	0,5



Poikkeava näyte sarjassa	Eron kuvailu N=3, (jos eron suuruus on > 0)
VRT N=2	Vähemmän hapan, Makeampi
KOE N=1	Ei niin keinotekoinen maku

GURT

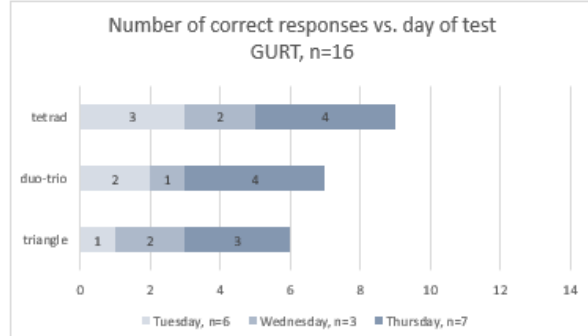
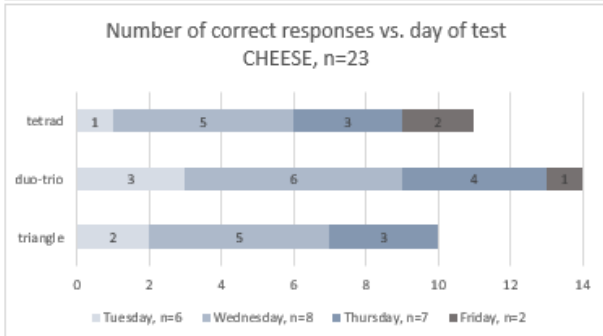
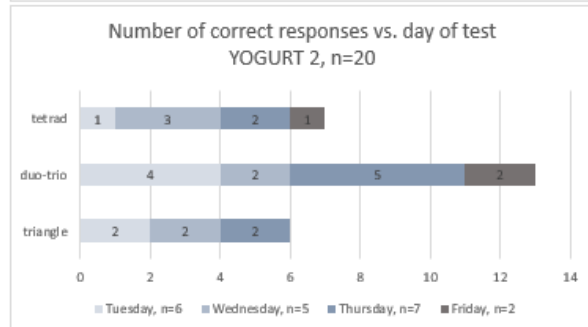
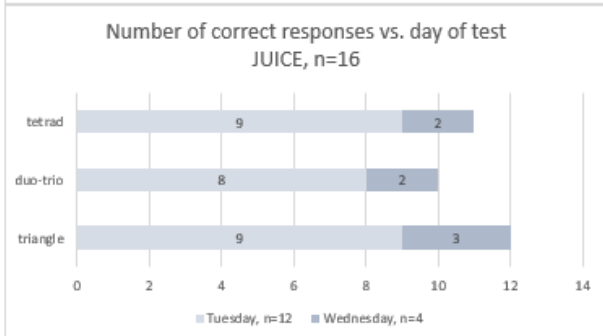
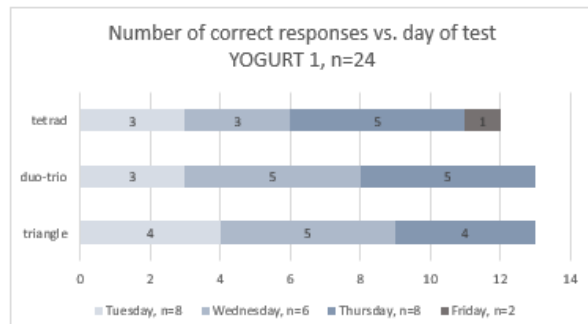
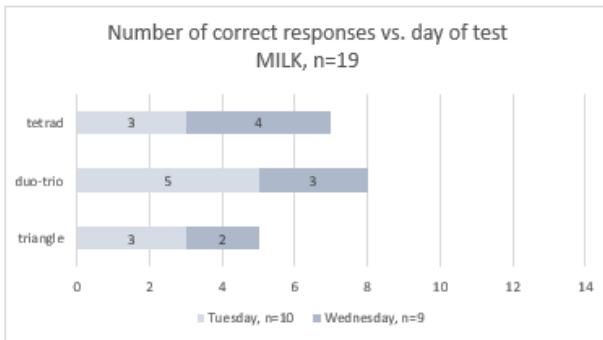
**ERON SUURUUS –
VAIN KOLMITESTISSÄ OIKEIN VASTANNEET, N=6**Eron suuruus vain eron havainneiden keskuudessa:

	VRT vs. KOE Frekvenssi N=6
0 = ei eroa	2
1 = pieni, juuri havaittava ero	2
2 = selvä ero	2
3 = erittäin selvä ero	0
keskiarvo	1,0

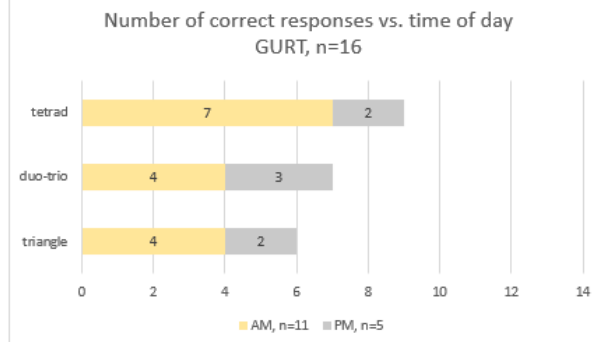
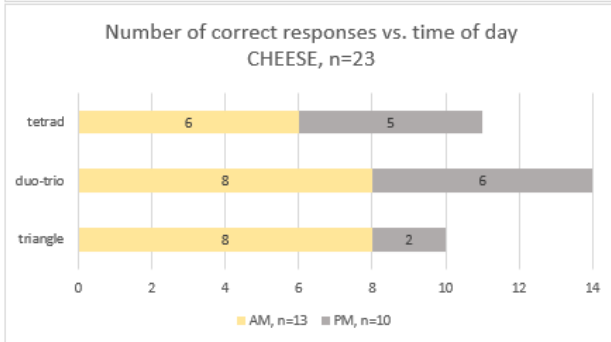
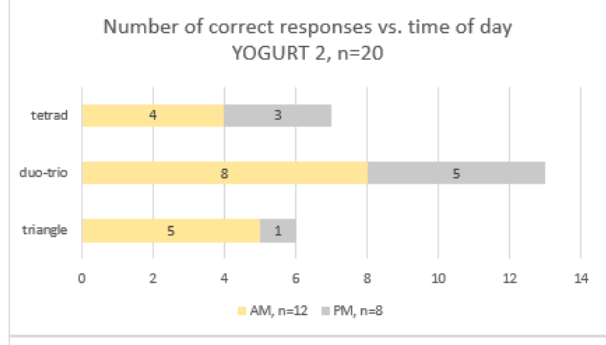
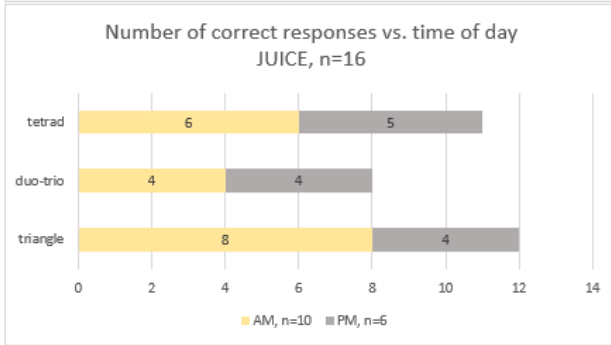
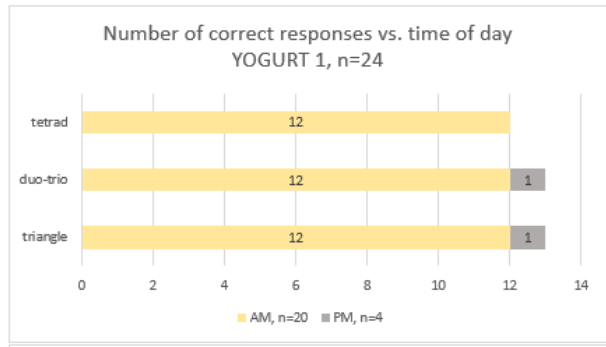
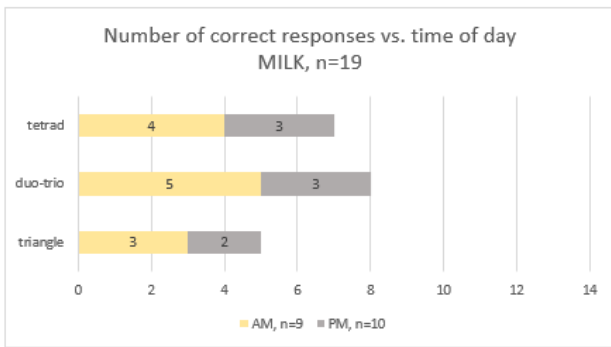


Poikkeava näyte sarjassa	Eron kuvailu N=4, (jos eron suuruus on > 0)
VRT N=2	vaaleampi väri massassa. eri maku ja rakenne.
KOE N=2	miedompi maku, vähän ohuempi rakenne, aromi, suuntuntuma ja jälkimaku

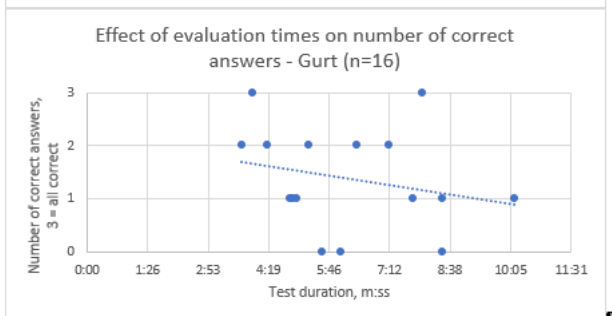
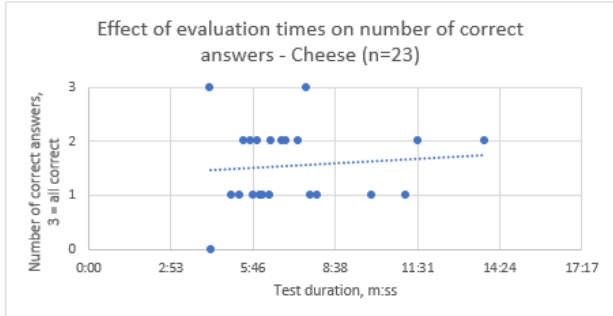
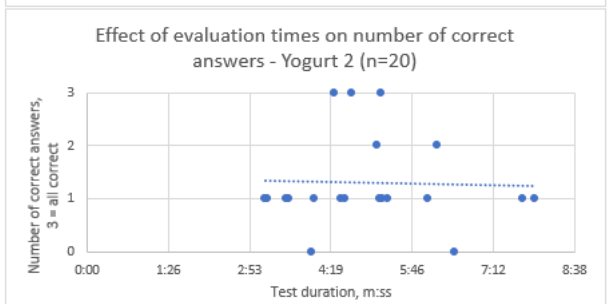
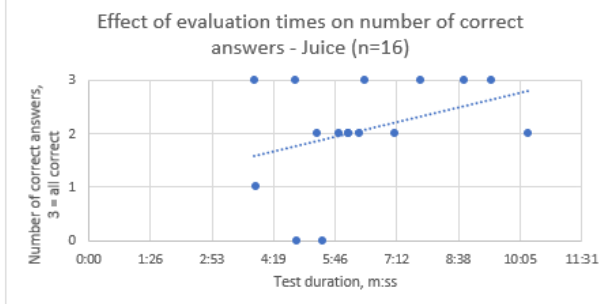
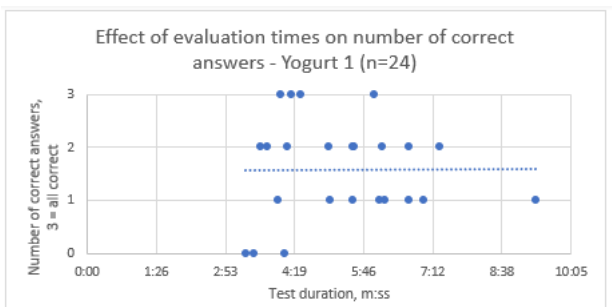
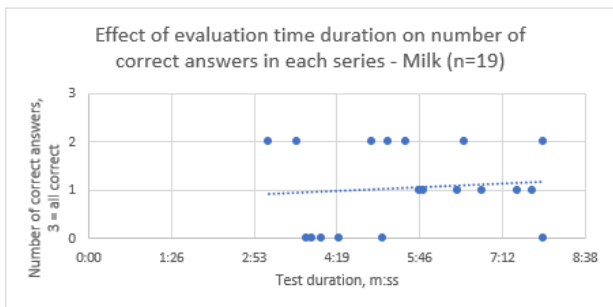
B: Number of total correct responses vs. day of test



C. Number of total correct responses vs. time of day



D. Number of individual's correct responses vs. time used in performing the test



Appendix 4. Overall feedback

OVERALL FEEDBACK

MILK

- *olipa vaikea*
- *vastasin 'oikein' kolmen näytteen sarjoissa, mutta neliotestissä 'väärin'. en tiedä joutuuko siitä, että neliotesti on viimeinen sarja ja kahden sarjan jälkeen on haastava jo löytää eroja.*
- *Mukavaa kun kolmitestin sijaan on vaihteeksi muitakin testejä. Itse en ole vakuuttunut kolmitestin kaikkivoipaisuudesta, koska maistaja ei tiedä mihin ominaisuuteen tulisi keskittyä (maku, makuvirhe, rakenne). Loppupeleissä moni kuluttaja saattaa kuitenkin huomata eron, vaikka tilastollista eroa ei tullut. Lisäksi tulostenkäsittelyssä minua hämää se että vaaditaan 95 % todennäköisyys: jos tulos on vaikkapa 80 % (=ei tilastollista eroa), niin eikö ero kuitenkin 80 % todennäköisyydellä havaita?*
- *Tämä oli hyvin mielenkiintoinen testaustapa. Toivottavasti jatkossakin on näitä. :)*
- -

YOGURT 1

- *Tosi hyvin ja turvallisesti hoidettu! Kiitos.*
- *näissä kaikissa oli enemmän eroa rakenteessa kuin maussa, toinen oli hieman paksumpaa kuin toinen.*

JUICE

- *kiva ja mielenkiintoinen testi. Kiitos*
- *näytteet liian kylmiä, vaikea arvioida*
- *Omaan makuun mehu on yllättävän makea, alkumaku tuntui jopa päärynäiseltä. Odotin että ero olisi ollut makeutuksessa näytteiden välillä, eipä ollutkaan :) Kun nyt tiedän että kyse oli raaka-aine-erosta, jonka ei tulisi pistää makunystyröihin, en usko että kuluttaja tätä huomaisi.*

CHEESE

- *n/a*
- *Tämänkaltainen testaussarja on todella mielenkiintoinen ja mukava! :)*
- *Luotin väriin enemmän kuin makuun eli juustoissa e ihan toimi:)*

YOGURT 2

- *Melkoisen rankka kokonaisuus maistaa näin monta erotustestiä yhdellä kerralla. Erot on kuitenkin pieniä.*

Appendix 5. Participant background

MILK

	Kaikki
N=	19
Sukupuoli	
Nainen	79 %
Mies	21 %
Muu	0 %
Ikä	
alle 25 v.	0 %
25-34 v.	16 %
35-44 v.	32 %
45-54 v.	26 %
55-64 v.	26 %
vli 65 v.	0 %

YOGURT

	YOGURT 1 Kaikki	YOGURT 2 Kaikki
N=	24	20
Sukupuoli		
Nainen	88 %	85 %
Mies	12 %	15 %
Muu	0 %	0 %
Ikä		
Yli 65 vuotta	4 %	0 %
55-64 vuotta	21 %	25 %
45-54 vuotta	21 %	25 %
35-44 vuotta	38 %	35 %
25-34 vuotta	17 %	15 %
Alle 25 vuotta	0 %	0 %

JUICE

	Kaikki
N=	16
Sukupuoli	
Nainen	81 %
Mies	19 %
Muu	0 %
Ikä	
Yli 65 vuotta	0 %
55-64 vuotta	12 %
45-54 vuotta	25 %
35-44 vuotta	38 %
25-34 vuotta	25 %
Alle 25 vuotta	0 %

GURT

	Kaikki
N=	16
Sukupuoli	
Nainen	88 %
Mies	12 %
Muu	0 %
Ikä	
Yli 65 vuotta	0 %
55-64 vuotta	25 %
45-54 vuotta	31 %
35-44 vuotta	44 %
25-34 vuotta	0 %
Alle 25 vuotta	0 %

CHEESE

	Kaikki
N=	23
Sukupuoli	
Nainen	91 %
Mies	9 %
Muu	0 %
Ikä	
Yli 65 vuotta	0 %
55-64 vuotta	17 %
45-54 vuotta	26 %
35-44 vuotta	35 %
25-34 vuotta	17 %
Alle 25 vuotta	4 %

Appendix 6. Sample tray pictures

MILK



YOGURT 1



YOGURT 2



JUICE



GURT



CHEESE

