Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

The Faculty of Natural Resources and Agricultural Sciences

Design and Performance of Small Scale Sensory Consumer Tests

Linn Svensson

Agronomist program - Food Science
Master's thesis • 30hec • Second cycle, A2E
Publikation/Sveriges lantbruksuniversitet, Institutionen för livsmedelsvetenskap, **no** 354
Uppsala 2012

Design and Performance of Small Scale Sensory Consumer Tests

Linn Svensson

Supervisor:	Cornelia Witthöft, Department of Food Science, SLU
Assistant Supervisor:	Lotta Beckman, Unilever Marie Rydén , Unilever
Examiner:	Monika Johansson, Department of Food Science, SLU

Credits: 30 hec
Level: Second cycle, A2E
Course title: Independent project/degree project in food science – Master's thesis
Course code: EX0425
Programme/education: Food Science Agronomist

Place of publication: Uppsala
Year of publication: 2012
Title of series: Publikation/Sveriges lantbruksuniversitet, Institutionen för livsmedelsvetenskap
no: 354
Online publication: http://stud.epsilon.slu.se

Key Words: Sensory evaluation, affective test, consumer test, preference test, acceptance test, the 9-point hedonic scale test, ranking test

Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

The Faculty of Natural Resources and Agricultural Sciences

Abstract

Small scale internal consumer tests provide a company with a cheap way to get valuable information regarding their products advantages and flaws. Therefore the demand for possibilities to do this kind of test has increased. This report is part of a new sensory project at Unilever. It presents a method for sensory comparison of the company's own products with the corresponding competitor products. The method allows the company to identify possibly poor performing products and also outstanding good performing products, why expensive big scale consumer tests can be limited into only including products that likely needs improvements or that in a larger scale test could be proven "best in test". Out of many potential tests for affective evaluation of foods the 9-point hedonic scale test and the ranking test was chosen. Questionnaires and preparation procedures were constructed, where after 16 evaluations including 23 Unilever products, performed on different food categories, were conducted. The results showed that indications and also significant differences in liking and preference could be seen in test groups of only ten participants. Among the evaluations performed seven Unilever products showed potential of being best in test, and six Unilever products got results indicating they were less preferred/liked compared to the competitor, why further evaluations are needed.

As an example of a possible way to proceed with the identified poor performing products, a second method, including the 9-point hedonic scale with added attributes and Just-About-Right scales, was presented and practiced on one product. The results showed that a product can be poor performing and graded as disliked among participants when compared to competitor products, but when tested on its own regarded as acceptable or even good performing. Further evaluations will have to be performed before determination if this is a successful method of identifying which attributes causes the product flaws.

Sammanfattning

Utförande av småskaliga interna konsumenttest är ett billigt sätt för företag att erhålla värdefull information angående fördelar och brister hos sina produkter. Av den här anledningen har efterfrågan efter nya möjligheter att utföra den här typen av undersökningar ökat. Den här rapporten är en del av ett nytt sensoriskt projekt på företaget Unilever. Rapporten presenterar en metod för sensorisk jämförelse av företagets egna produkter med motsvarande konkurrenters produkter. Metoden gör det möjligt för företaget att identifiera de produkter som tenderar att prestera sämre än konkurrenternas, samt de produkter som presterar enastående bra. På detta sätt kan utförandet av dyra storskaliga konsumenttest begränsas till att endast inkludera de produkter vilka sannolikt är i behov av förbättringar eller vilka i ett större test skulle kunna påvisa ”bäst i test”. Av flera potentiella test för affektiv bedömning av livsmedel valdes den 9-gradiga hedoniska skalan samt ranknings test ut. Frågeformulär samt förberedelseförfaranden konstruerades, varefter 16st utvärderingar/tester, innefattande 23st Unileverprodukter, genomfördes på ett antal olika livsmedelskategorier. Resultatet visade att det är möjligt att se indikationer och även signifikanta skillnader i tycke och preferens i testgrupper innehållande endast tio stycken deltagare. Bland undersökningarna/testerna som utfördes visade sig sju Unileverprodukter ha potential att vara ”bäst i test”, medans resultaten för sex Unileverprodukter indikerade ett sämre resultat i jämförelse med konkurrenter varvid det behöver göras vidare utvärderingar och tester.

Som ett exempel på hur man skulle kunna gå vidare med de produkter som identifierats som eventuellt sämre i jämförelse med konkurrenter konstruerades en andra metod, vilken genomfördes på en av produkterna. Metoden innefattade den 9-gradiga hedoniska skalan med specificerade attribut samt ”Just-About-Right”-skalor. Resultaten visade att en produkt kan vara sämre/mindre omtyckt i jämförelse med konkurrerade produkter, men då den testas separat anses acceptabel eller till och med bra. Vidare undersökningar måste genomföras för att avgöra om metoden är framgångsrik vid identifiering av vilka attribut som orsakar en produkts brister.

Table of contents

1. Introduction

During the second half of the twentieth century the field of sensory evaluation has grown rapidly. Part of the reason to the rapid growth is the expansion of the processed foods and consumer products industries (Lawless and Heymann, 2010). A commonly used and accepted definition for sensory evaluation, introduced by the (U. S) Institute of Food Technologists in 1975 (Dijksterhuis, 1997) (Lawless and Heymann, 2010) (Stone and Sidel, 1993), is;

"A scientific discipline used to evoke, measure, analyze, and interpret those reactions to the characteristics of products as they are perceived through the senses of sight, smell, touch, taste and hearing"

An article by Koehl et al. (2007), adds another clear and simplified description of sensory evaluation;

"*Under predefined conditions, a group of organized individuals evaluate attributes of a group of products with respect to a given target"*

Sensory evaluation methodologies attempt to isolate the sensory properties of foods to enable accurate measurements of the human responses. This is done by minimizing potential biasing effects, like information that could influence consumer perception, for example brand image and price etc. (Lawless and Heymann, 2010).

In today's companies all activities and decisions are coloured by the current consumer preferences (Dijksterhuis, 1997). Through different kinds of sensory evaluations, companies can get important and useful information regarding both sensory characteristics of their products and information regarding the consumer liking and preferences (Lawless and Heymann, 2010). This information is crucial in determining and maintaining the quality of a product, in the work towards new product development, in the forecasting of market behaviour and when exploiting new markets (Koehl et al., 2007).

1.1 Aims and purpose of the project

The aim of the project was to develop and implement methodology that can be used for small scale sensory evaluation of food products, to see how they perform in a sensory point of view compared to competitor products. In the first step the purpose was to find a method that can identify products that are not reaching good enough criteria. In a second step the purpose was to find a method that further investigates the reason behind the result, and identifies what attributes that causes the bad outcome. At last some suggestions on actions are presented.

1.2 Project description and delimitations

This report consists of a literature review covering different sensory evaluation methods and common practices when performing tastings of food products. In addition to the literature study, construction of suitable evaluation questionnaires and practical work regarding evaluation of a number of food products was performed. When structuring the project, the following questions have been regarded;

- What kinds of sensory evaluation methodology are there, which of them are suitable for the project, and how can they be implemented to suit the present conditions at the company?

- How to develop sensory evaluation questionnaires that answer the questions concerned and are easy to use for the company.

- What are the common practices when performing a sensory evaluation session, and how can these be implemented at the company?

- How to analyze the collected data to discover how the chosen Unilever food products perform in comparison to competitor products.

- How can the data collected from the sensory evaluations be of help for the company? What actions can be taken to make improvements?

The literature research has been done through databases; ScienceDirect and Scopus, the SLU library and the Uppsala library. For more detailed information regarding the literature and material used in the report see references at page 39-40.

To delimit the project the following areas have been excluded:

- This study does not in depth evaluate the differences i.e. Pros and Cons, of different affective sensory evaluation techniques/tests.

- This report serves as a first step in the start up of a new sensory evaluation project at Unilever. The methodologies and questionnaires used might therefore be altered along with their implementation and gathering of experience.

- This study mainly focuses on competitor tests, why methodology for finding the reasons behind a bad performing product will be brief.

2. Sensory evaluation methodology/techniques

Sensory evaluation techniques are often divided into three categories/classes: discrimination analysis, descriptive analysis and affective analysis. The division is based on the goal of the study i.e. the question to be answered, and on the criteria/characteristics demanded for the participating panellists.

1) Discrimination analysis answer the question: "are the products different in any way?"
2) Descriptive analysis answer the question: "how do products differ in specific sensory characteristics?"
3) Affective analysis gives answers to: "how well are the product liked or which product is preferred? "

Discrimination tests and descriptive tests are both analytical tests, but whereas the descriptive tests require a trained panel, the participants in a discrimination test just needs to be partly trained or sometimes do not need training at all. However, when performing affective tests only untrained panels should be used (Lawless and Heymann, 2010). O'Mahony (1998) divided sensory evaluation into two types only. In Type I trained panellists are used, and reliability and sensitivity are the key factors. In type II the participants should represent the consuming population, and the evaluation is supposed to be more naturalistic (Lawless and Heymann, 2010). Nevertheless, in this report we will use the three classes division.

2.1. Discrimination analysis

Discrimination analysis comprises the simplest tests for sensory evaluation. However, they have proved to be very useful and are heavily used (Lawless and Heymann, 2010). Discrimination tests serves to discover if there is any perceptible difference among samples/products, and they are often divided into two groups:

1) Overall difference tests
2) Attribute difference tests

In the overall difference tests participants are asked if they can perceive any existing difference at all between samples, and in the attribute difference test participants are asked to focus on a specific attribute; for example rank the samples after degree of sweetness (Meilgaard et al., 2007). Discrimination analysis can be very useful in product development when investigating new possibilities in reformulating a product recipe or processing without creating a detectable change for the consumer. A company might want to switch an expensive ingredient into a less expensive one, or want to change some steps in the processing of the product. It is then of great interest for the company to be absolutely sure that the consumer will not perceive any difference between the old and the new version (Lawless and Heymann, 2010). Another scenario is when the food company wants to create a "new and improved" version of an already existing product. In this case they want to detect a difference between the old and the new version and be sure that the consumer also will perceive it (Lawless and Heymann, 2010).

The participants in a discrimination test do not need any heavy training, but should preferably be familiar with the test procedure and have been screened for sensory acuity. An adequate sample size, to be able to document clear sensory differences, when performing discrimination tests is 25-40 participants (Lawless and Heymann, 2010). Nevertheless, some

discrimination tests can be performed with as few as six participants if differences between samples are large (Meilgaard et al., 2007).

A number of different discrimination tests exist. Some of the most commonly used are: the triangle test, the duo-trio test and the paired comparison test (Lawless and Heymann, 2010). The triangle test and the duo-trio test belong to the overall difference tests, while the paired comparison test is an example of an attribute difference test (Meilgaard et al., 2007). In the triangle test the participants are asked to, among three samples, choose the sample that is most different. In the duo-trio test the participants are asked to point out the sample that matches the given reference sample. In the paired comparison test the participants are asked to tell which sample that is most intense in a given attribute, e.g. the sweetest (Lawless and Heymann, 2010).

2.2. Descriptive analysis

Among the sensory evaluation tools, the descriptive analysis (DA) have shown to be the most informative and comprehensive, giving a lot of detailed information (Lawless and Heymann, 2010). Descriptive analysis comprises detection and description of both qualitative and quantitative aspects of a product, i.e. to localize characteristic attributes and to quantify the perceived intensities of the sensory characteristics of a product (Meilgaard et al., 2007). Descriptive analysis can for example be used when wanting to know which attributes that have changed in sensory characteristics of a new product (Lawless and Heymann, 2010).

The methods used in descriptive analysis often consist of developing and using a list of sensory attributes and their intensity. The list of attributes in combination with the scales of intensity gives an evaluation questionnaire that is useful in measuring the differences among samples. In descriptive analysis, the evaluation questionnaire almost always needs to be uniquely constructed to suite the product and the question to be answered (Lundgren, 2000). In descriptive analysis a well trained panel is always used. The selection of the individuals to the panel is based on having average to good sensory acuity for the important characteristics; taste, smell, texture etc. (Lawless and Heymann, 2010) (Meilgaard et al., 2007). Through training the panellists learn to adopt an analytical frame of mind where they focus on specific product aspects. Together the panellists get calibrated into using scales on the questionnaires in an analogous way. When performing descriptive analysis the panellists must put their personal preferences aside and work as an analytical instrument; focusing on specifying what attributes that are present and at what level/extent. A trained panel can easy agree in the use of words describing the attributes of a product, while consumers often differ in great extent when transforming impressions to words (Lawless and Heymann, 2010). In most cases an adequate sample size when performing descriptive analysis is about 8-12 participants, however larger panels with up to 100 participants can be used when evaluating products for mass production where small differences can be of great importance (Meilgaard et al., 2007).

Descriptive analysis has been proven to be a very useful tool in product development due to that it can be used to characterize a wide variety of product changes and give a detailed specification of a product's sensory attributes (Lawless and Heymann, 2010). Through descriptive analysis it is possible for the company to see exactly how the own product and the competitor product differ in the sensory dimension. It is also a great tool when testing product shelf-life or when wanting to define a sensory problem to be able to improve the quality. The information collected through descriptive analysis can also often be related to information regarding consumer preference. Since descriptive techniques tend to be expensive, they are

commonly not used in the day-to-day quality control, but mostly used when troubled with major consumer complains (Lawless and Heymann, 2010). Examples of descriptive tests are; The Spectrum Descriptive Analysis Method (Meilgaard et al., 2007), The Profile Attribute Analysis Test, The Texture Profile Test, and The Sensory Spectrum Procedure (Lawless and Heymann, 2010).

2.3. Affective analysis

The affective analysis, also called acceptance tests, preference tests or hedonic tests, are used to quantify the consumer preference or degree of liking/disliking of a product (Lawless and Claassen, 1993). The purpose is to evaluate the personal response of preference or acceptance from current or potential customers concerning a product idea, an existing product or some specific product characteristics (Meilgaard et al., 2007). Affective tests are so-called consumer test which means that the participants in the study always should be untrained and representatives of the consuming population. It is not wise to let trained panellists answer questions regarding preference and liking since they have a too analytical way of evaluating. Consumer often react immediate and perceive the product as a whole pattern, without considering different attributes in detail or putting a great deal of thought into the evaluation. This integrated way of evaluating a product is expressed in liking or disliking of the product (Lawless and Heymann, 2010). The participants in an affective test should be regular users of the product i.e. belong to the target market, or at a minimum like the type of product that is tested and be familiar with similar products. By choosing participants within these criteria it is made sure that the participants have a frame of reference and thereby can compare the product with similar products that they have tried. It also makes sure that the participants possess reasonable expectations on the product (Lawless and Heymann, 2010).

When performing affective tests, an adequate sample size is around 75-150 individuals (Lawless and Heymann, 2010), or even larger; 100-500 individuals (Meilgaard et al., 2007). The reason to why the panel size in an affective test needs to be large is that individual preference has such a high variability (Lawless and Heymann, 2010). The preference of individuals can differ in many different ways, because of several different reasons, e.g.: Personal background, experiences, culture, attitudes and habits. Personal interests, like interest in health, believe in different diets, interest in environment, practicing of sports etc. can also affect personal preference. All of these reasons affect each individual's preference regarding e.g. appearance, texture, smell and taste of a food. In addition to this, individual's likes and dislikes may be affected due to: the time of the day for consumption, the number of times they have consumed the food recently, the serving temperature of the food and if this is consistent with the individual's expectations (Lawless and Heymann, 2010). To get enough sensitivity and statistically power in an affective test the sample-size therefore needs to be increased.

Affective tests give opportunities to find segments of people who prefer different styles of the products, and can also lead to discovery of the reasons behind why the different segments of people having certain preferences (Lawless and Heymann, 2010). The use of consumer tests have become more common in recent years, as they have proven to be a highly effective tool in predicting consumer preferences to be able to develop and produce products that will sell in large quantities or allows a higher prices (Meilgaard et al., 2007). Affective tests are often used as a part of market consumer surveys, to get the consumer perspective/opinion on products i.e. localize product benefits and product flaws (Earle et al., 2001). Generally the reason behind the company wanting to conduct consumer tests are; product maintenance,

product improvement or optimization, development of new products, assessment of market potential, product category review, and support for advertising claims (Meilgaard et al., 2007).

With the large panel size follows a need for a larger amount of product samples. This in addition to the need of great amounts of time for preparations and implementation can make affective tests become very expensive. Due to this fact, alternative approaches and new cost-effective options are constantly developed. There is both qualitative and quantitative testing and they include; in-house panels, home use test, focus groups and online research. To be cost-effective many companies today perform in-house product screening prior to larger tests in market research (Meilgaard et al., 2007).

3. Sensory evaluation techniques suitable for the project

Depending on the objectives of an investigation the category of test method and best specific tool for sensory evaluation must be chosen. In a project containing multiple objectives a sequence of different tests sometimes can be required (Lawless and Heymann, 2010).

In this project we wanted to find a method that answers the question; how does the company's own product perform, from a sensory point of view, in comparison with the competitor's corresponding product? Since this is a question that regards preference and liking, an affective test is the proper choice. To get to know which product that is most liked or most preferred an appropriate affective test needs to be chosen, and a consumer study be performed. The affective tests are sometimes divided into preference tests and acceptance tests (Lawless and Heymann, 2010). There are several different tests to choose from. A short review of some of the most commonly used tests are presented below.

3.1. Acceptance test

Acceptance tests measure the degree of liking or disliking by the use of rating scales. Examples of acceptance tests are; the 9-Point Hedonic Scale, The Labeled Affective Magnitude Scale (LAM), Line Scales, Just-About-Right scales (JAR) and Food Action Rating Scale (FACT) (Lawless and Heymann, 2010).

3.1.1. The 9-point hedonic scale

The hedonic rating scales are used to quantify affective dimension of the consumer perception of foods (Tuorila, 2008). Among the hedonic rating scales, the 9-point degree of liking scale, also called the 9-point hedonic scale, is probably the most commonly used (Tuorila, 2008) (Lawless and Heymann, 2010). The scale was invented in the 1940s and has been carefully developed, tested and evaluated during the years (Lawless and Heymann, 2010). In the test participants/consumers are asked to give their hedonic opinion to a product sample by choosing and marking one of nine alternatives, (ranging from 1 = like extremely to 9 = dislike extremely). The 9-point hedonic scale is nowadays present in several different appearances (Lawless and Heymann, 2010). The verbally anchored scale is probably one of the most used forms (Tuorila, 2008). Here every option on the 9-point scale has a verbal expression. Behind the verbal anchors/expressions lies comprehensive work and research to ensure and validate that each scale option is based on almost equal differences to give the scale ruler-like properties. However, the precise construction of the 9-point hedonic scale and

its verbal anchors limits the use of the scale in other languages than English (Lawless and Heymann, 2010). The 9-point hedonic scale can be seen printed both vertically and horizontally, and with the like side or the dislike side first. The way the scale is printed on the score sheet should, according to studies made by the Quartermaster Institute, not affect the results (Jones et al., 1995). Except for the verbal anchored scale, the hedonic 9-point scale also exists in different modifications. For example can the scale be used without the verbal labels, or it can be altered into becoming unbalanced, containing more like than dislike options (Lawless and Heymann, 2010).

Figure 1. The Figure shows the verbally anchored 9-point hedonic scale/"The degree of liking scale".

The ruler-like properties of the 9-point hedonic scale gives it some advantages that other less carefully constructed liking scales often not possess. It makes it possible to assign numerical values to the scale and use parametric statistics when analyzing the data. Other advantages with the 9-point hedonic scale are that it is easy to use and clear and easy to understand for the participants. It has shown to be reliable, possessing high stability of responses and is to some degree flexible regarding panel size (Lawless and Heymann, 2010).

Among negative criticism is that the interval spacing sometimes, when comparing results with direct scaling methods, has been accused for not being totally equal in terms of distance between "neither like nor dislike" and "like slightly"/"dislike slightly". It has been said that this distance is smaller than the other intervals. However initial calibration work of the scale has shown the spacing to be equal. Other discussed criticism is the risk of "end use avoidance" i.e. when participants avoid the extreme categories. However this is a problem that is not unique for the 9-point hedonic scale (Lawless and Heymann, 2010).

3.1.2. The Labeled Affective Magnitude Scale (LAM)

LAM is a modification of the 9-point hedonic scale. It was developed in 2001 as an alternative when measuring food acceptability and consumer liking. LAM is based on magnitude estimation, which means that the participants in the tests get to use any number they want to describe their liking, but should focus on the ratios/proportions between the products. The ratios between the numbers are meant to reflect the ratios of experienced sensation magnitude. For example if participants give product A the value 30 for preference/liking and then perceive product B as twice as liked/preferred, product B should get a value of 60. Except for the rescaling into using magnitude estimation, the LAM line scale also got added anchors; "greatest imaginable like" and "greatest imaginable dislike" (Lawless and Heymann, 2010).

Some studies of performance of LAM scales compared to the 9-point hedonic scale has, when comparing well liked foods, shown that the LAM scale sometimes is better than the 9-point hedonic scale (El Dine and Olabi, 2009), however studies done by Schutz and Cardello (2011) have shown them to be similar in performance. Another study made by Lawless et al. (2009), where several different food categories were evaluated, showed that in some cases LAM was found to be best and sometimes the 9-point hedonic scale was superior.

3.1.3. Line scales

Line scales, also called visual analogue scales (VAS) are the standard scaling method in descriptive analysis. However, they are sometimes also used in affective analysis (Lawless and Heymann, 2010). Line scales are unstructured scales that in hedonic tests often are anchored with like and dislike in the ends and sometimes also a middle point for "neither like nor dislike". Line scales can differ slightly from one another by either be marked or unmarked (*Figure 2*). The anchors in the end points can also be expressed in slightly different ways (Lawless and Heymann, 2010). When analyzing the results, the marks on the line scales are converted into numbers by the use of a ruler or a computer (Meilgaard et al., 2007). Marked Line scales have in tests shown to have an advantage over the 9-point hedonic scale in terms of product differentiation and identification of consumer segments (Villanueva and Da Silva, 2009).

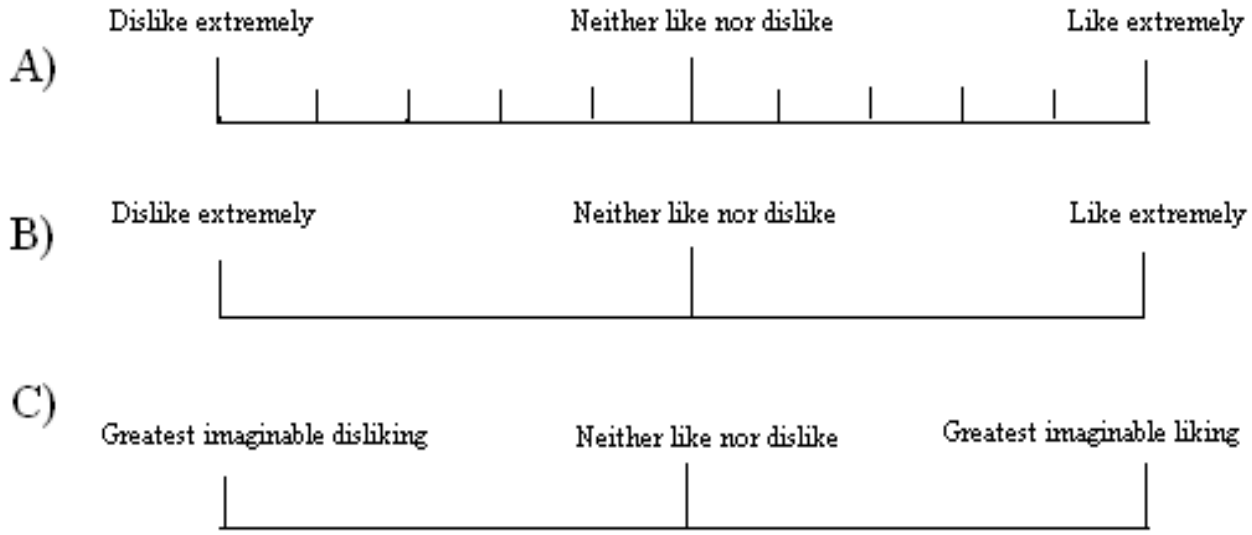

Figure 2. The figure shows line scales for acceptability testing. A) Marked line scale, B) Unmarked line scale, C) Simplified LAM scale. (Inspiration from Lawless and Heymann, 2010)

3.1.4. Food Action Rating Scale (FACT)

Food Action Rating Scales are based on statements regarding frequency of consumption and motivationally related statements (Lawless and Heymann, 2010). The number of statements and the formulation of the sentences reflecting the affective action can vary (*Figure* 3). According to a study made by Howard G. Schultz, 1965, where the FACT-scale was compared with the 9-point hedonic scale, the FACT-scale was shown to be a reliable and sensitive method for evaluating food acceptance. In the study the FACT-scale even showed to be more sensitive than the hedonic scale, which was interpreted as a sign of the scale being

even easier to use by participants compared to the hedonic scale. However, one big limitation in the usage of FACT-scales is that the FACT-scale only can be used as an over-all measure of food acceptance i.e. it can not give any information regarding specific attributes like texture, appearance, aroma etc. (Schutz, 1965).

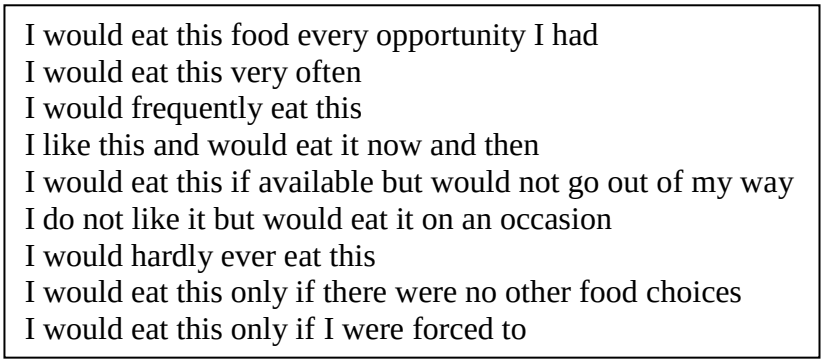

Figure 3. The figure shows an example of how a Food Action Rating Scale (FACT-scale) can be designed (Lawless and Heymann, 2010).

3.1.5. Just about right scales (JAR)

Just about Right scales measure the consumer's reaction to a specific attribute, for example sweetness, saltiness, crunchiness, thin/thickness etc. (Lawless and Heymann, 2010). They are among others often used in food research and development to determine the optimal level of a specific ingredient (López Osorino and Hough, 2010). The JAR-scales are bipolar, and the end anchors are always true opposites like; "very much too thick" and "very much too thin", or "very much too salt" and "very much not salt enough". The centre point represents the point where the regarded attribute is just right, and is therefore labelled "just right" or "just about right" (Lawless and Heymann, 2010).

The Just About Right Scale combines intensity and hedonic judgment (Rothman and Parker, 2009) and can give directional information for product reformulation by testing only one single product. The fact that Just about right scales gives direct information on specific attributes has made them very popular and they have proven to be a good tool in product development (Lawless and Heymann, 2010). Through JAR scales it is possible to get diagnostic or explanatory information if the overall product appeal is lacking (Lawless and Heymann, 2010).

A positive aspect of JAR is the simplicity of the test. However, there are also negative aspects to consider. Only attributes having an optimum can be used i.e. attributes where more (or less) always is better are not suitable, and the attribute must not have a negative association. There is always a risk that the participant might misinterpret the attribute, or that the test asks for a too analytical way of evaluation the food than the consumer can handle (Lawless and Heymann, 2010).

Saltiness	**Thickness**
□ Very much too salt	□ Very much too thick
□ Too salt	□ Too thick
□ Slightly too salt	□ Slightly too thick
□ Just about right	□ Just about right
□ Slightly not salt enough	□ Slightly too thin
□ Not salt enough	□ Too thin
□ Very much not salt enough	□ Very much too thin

Figure 4. The figure gives an example of two ways of constructing a JAR-scale. (Inspiration gathered from Lawless and Heymann, 2010).

3.2. Preference tests

In a preference tests the consumer is asked to tell which of two or more samples that he/she prefers. Examples on commonly used preference tests are: Paired Preference Testing, Ranking Test and Best-Worst Scaling (Lawless and Heymann, 2010).

3.2.1. Paired Preference Test

When performing a paired preference test two samples are presented to the participant simultaneously, and the participant is asked to evaluate the samples and tell which one he/she prefers. In most cases the participant is forced to make a choice between the samples, however sometimes a no preference option is included (Lawless and Heymann, 2010). One limitation with the paired preference test is that it does not indicate whether any of the two products tasted are liked or disliked or to what degree one of the products are more preferred over the other. The paired preference test is therefore most suitable for products where there is a prior knowledge of the "affective status" (Meilgaard et al., 2007).

3.2.2. Ranking Test

In a ranking test the participant is asked to rank a number of products in descending or ascending order according to preference or liking (Lawless and Heymann, 2010). It is used when the objective of the test is to compare several samples according to one single attribute, for example; overall preference, freshness, saltiness etc. (Meilgaard et al., 2007). The ranking test has several similarities with the paired preference test (earlier described) e.g. it most often uses forced choice (Lawless and Heymann, 2010), it gives ordinal data and gives no indication of degree of difference between samples (Meilgaard et al., 2007).

3.2.3. Best-Worst-Scaling (BWS)

Best-Worst scaling (BWS), sometimes called "maximum difference scaling" or "maxdiff", is quite similar to the raking test and the paired comparison test described earlier, however here the participant should only pick out which product he/she prefers/like the most and the product that he/she prefers/likes the least. Just like in the ranking test and in the paired comparison test BWS does not give any answers on the magnitude of liking/preference and the magnitude of difference in preference between samples. Best-Worst scaling is used when more than two products are compared (Lawless and Heymann, 2010). By calculating the

number of times a product is voted as best and subtract this value with the number of times the product is voted as worst, a value is gotten that makes the products in the test easy to compare (Jaeger et al., 2008). According to a study done by Jaeger et al., 2008, usage of BWS is perceived as easy to use and understand by the participants in the test.

4. Common practices when performing sensory evaluation tests

When performing sensory evaluation tests there are many factors that needs to be taken into consideration and to be controlled if the results are to be reliable. Planning the tasting session carefully is of great importance. In the preparation practices there are three major areas to consider; test controls, product controls and panel controls. Test controls concerns the environment in which the tasting is performed i.e. the layout of the tasting room and the preparation area etc. The product controls regards equipment, the preparation, numbering, coding and serving of the samples etc. Panel controls comprehends the procedures performed by the participants in the panel when evaluating the samples (Meilgaard et al., 2007).

4.1. Test controls

When deciding on the location of the test area, there are some practical aspects to consider. The location should be easy to reach for the participants and free from crowding. Eliminating surrounding variables not originating from the products themselves minimizes participant's biases and maximizes their sensitivity. The environment should thereby be as free as possible from factors that could confuse or distract the participants, and be designed in a way that maximize their focus and sensitivity to the products tested. Therefore it is among others of great importance that the test area is free from odours and noise. Other aspects are the colour and lightning, air conditioning, relative humidity and temperature. Depending on what kind of sensory evaluation technique that is to be used the optimal layout of the test area may vary. If using methodology that builds on the participant's individual opinion, interaction between the subjects must be eliminated and in this case a test room including individual testing booths is to prefer. On the other hand, if performing a sensory evaluation where the participants are supposed to interact and come to consensus, a round table is more convenient. In connection to the test area there need to be a preparation and a storage area where samples can be prepared and stored. Depending on what products that are going to be tested the equipment needed in this area may vary. Usually the preparation area is similar to a kitchen i.e. there need to be cooking equipment, benches for preparation, refrigerator, freezer etc. There are many different layouts of testing and preparation areas, and there is literature that in detail describes how to shape the perfect facility for specific evaluation conditions. However, many companies that perform in house sensory evaluation do not have access to these perfect conditions and thereby have to construct the test area in the best possible way with the conditions given (Meilgaard et al., 2007).

4.2. Product controls

It is important that the equipment, early handling, preparation, and the presentation of the products are performed in a structured and controlled way. When choosing the equipment for preparation and presentation of the samples it is important to make sure that the materials must not transfer any volatiles (odours or flavours) to the product. It is important that the size of the serving is the same for all samples. When performing a consumer

preference/acceptance test the products should be served in the same way as they are normally consumed and according to the consumer's preference i.e. tea together with milk or/and sugar, soup together with bread, margarine/butter together with bread etc. However, when performing difference and descriptive tastings with trained panels the product is served on its own without any additives. The serving temperature must be appropriate why developing standard preparation procedures is crucial, or alternatively check of each sample must be done before serving. The order, coding and number of samples also must be monitored. The order in which the samples are presented should be balanced and each sample should appear in a position equal numbers of time. The presentation should be random. Single and double letters or digits are to be avoided since people might have favourite numbers or letters (Meilgaard et al., 2007).

4.3. Panel controls

Participants in a consumer test must get careful instructions regarding what is asked for in the test, i.e. what kind of judgment/evaluation that is to be made (preference, acceptance, description or difference), the handling of the samples and the use of questionnaires. Prior to the test they should be informed regarding the number of samples that are to be tested, the delivery system of the samples and on how to evaluate the samples and use the scales for expressing their judgment. To prepare the participants for the task and what they are to expect minimizes the risk of participants feeling uncomfortable, anxious or becoming distracted, which in turn minimizes the variation in the test design and extraneous variables that could bias the result (Meilgaard et al., 2007).

Depending on the product tested, selection of the participating individuals for a consumer test can be based on different demographic factors, like: age, gender, national origin, education level, income, culture, marital status, family size etc. The basic rule when assembling the panel for a consumer test i.e. deciding on the criteria for participation is that the participants should reflect the target market (Meilgaard et al., 2007).

5. Internal consumer tests

Internal consumer tests are tests conducted at the company, using employees as participants/panellists (Lawless and Heymann, 2010). Using employees as participants in consumer tests are often regarded as a real advantage by the company, since it can reduce the high costs that are associated with consumer tests. Employees are often familiar with the products, their characteristics and with the testing procedure and are therefore able to handle a larger number of samples and can often give faster answers which reduce the time of the test. The employees can thereby be seen as a valuable resource for the company giving the company cheap service in the consumer testing (Meilgaard et al., 2007). There are however some negative aspects related to internal consumer tests that are worth to be considered. Using employees as participants in a consumer study makes it impossible to sample properly from the consuming population (Meilgaard et al., 2007). There is a major risk in that employee might show liability towards having biasing information and assumptions regarding the products tested. Behind their choice in work lies most likely an interest in food greater than the average consumer's, and knowledge that might influence their assessment of a food (Resureccion, 1998). A food technologist might for example focus on entirely different attributes of the product than the average consumer.

Depending on the project objective it can be more or less suitable to use employees as participants in consumer tests. Generally employees are regarded as less of a risk as a test group if the consumer test regards product maintenance. In consumer tests regarding new-product development, improvement or optimization employees should not be used as participants in the tests (Meilgaard et al., 2007).

6. Material and methods

6.1. Evaluated products

Evaluations were performed in Sweden and in Finland. The number of food products/brands evaluated at each test occasion varied from two to five, and the number of individuals participating in the tests varied from 8-25. Three different food categories were investigated; margarines, savoury: soups, sauces and bouillon, and tea. Following 16 evaluations of 23 Unilever brands compared to the corresponding products of competitor brands were performed (these evaluations will be referred to as "competitor tests/evaluations"):

Margarines:

- Margarine 1 (three brands) - Sweden
- Margarine 2 (five brands) - Sweden
- Margarine 3 (five brands) - Sweden
- Margarine 4 (four brands) - Finland
- Margarine 5 (three brands) - Finland

Savoury:

- Soup 1 (three brands) - Sweden
- Soup 2 (three brands) - Sweden
- Sauce 1 - dry mix (three brands) – Sweden
- Bouillon 1 (two brands) - Finland

Tea:

- Tea 1 (four brands) - Sweden
- Tea 2 (five brands) – Sweden
- Tea 3 (two brands) – Sweden (NOTE: No ranking test performed)
- Tea 4 (two brands) - Finland
- Tea 5 (two brands) - Finland
- Tea 6 (three brands) - Finland
- Tea 7 (two brands) – Finland

Soup 1 has been further evaluated by the use of an additional test method, as an example of a possible way to proceed with poor performing products, to evaluate the reason behind the result and identify possible ways of improvement.

6.2. Sensory evaluation methods

The sensory evaluation methods used in the competitor tests were:

- The 9-point Hedonic Scale
- The Ranking Test

The sensory evaluation methods used for further evaluation of poor performing products were:

- The 9-point Hedonic scale with added attributes
- Just-About-Right Scales (JAR-scales)

The 9-point hedonic scale together with the ranking test was used for competitor testing of all product types, except for the evaluation of Tea 3, where the ranking test defaulted.

6.3. Adaption of sensory evaluation techniques and practices to suit the conditions at the company

Performing major consumer tests is very expensive, and in this project the aim was to perform these kinds of tests as quality control, to maintain a high standard and confirm that Unilever products are better or at least as good as their competitor's. The sensory evaluations should therefore be used as a routine measure, which means that they will be done frequently and on many products. This leaves no room for expensive and time consuming big scale consumer tests, why smaller scale internal consumer test (in-house-screening) was more suitable. Since the objectives of the performance of consumer tests was close to product maintenance the usage of internal customer tests with employees as participants was regarded as acceptable, despite some negative aspects (see 5, "Internal consumer tests").

For test- and preparation area the Unilever local café and diner for employees with connecting kitchen was utilized. In absence of testing booths each participant was seated alone at a separate table.

6.4. Methods used in analyzing of the results

When analyzing the results from all three sensory evaluation methods, histograms were drawn for all the products and sum, mean value, median value and standard deviation calculated (See example in appendix 2 and 3).

For six of the competitor tests: Margarine 3, Margarine 5, Tea 3, Tea 4, Tea 6 and Tea 7, it appeared that not all participants liked the food of the test, or had not verified that they liked the food of the test. Two calculations were therefore made: one calculation including only those participants that had verified they like the food of the tasting, and one calculation including all the participants.

In the summarizing of the results from the competitor tests, i.e. the 9-point hedonic scale test and the ranking test, four tables were drawn. The division into the tables was based on the sums of the products. All Unilever products having a better sum value i.e. a lower value, compared to the competitor were placed in *Table 1*. All Unilever products having a worse sum value, i.e. a higher value, compared to the competitor were placed in *Table 2*. This gives; In *Table 1* all Unilever products that potentially could be significantly more liked/preferred compared to the competitor brand are presented and in *Table 2* all Unilever brands that possibly could be significantly less preferred compared to the competitor brand are presented. Based on the same division, *Table 3 and 4* presents the results from the six competitor tests were double calculations were made.

6.5. Statistical methods

In the competitor tests the 9-point hedonic scale test was analyzed for significance, using parametric statistics; the Paired (Dependent Sample) t-test, and the ranking test was analyzed for significance by the use of non parametric statistics; the Friedman test in combination with calculations of "the Least Significant Ranked Difference" (LSRD). For the paired t-test, significance values ranging from 0.2 to 0.001 was used and registered as significant, and for the Friedman test significance values ranging from 0.1 to 0.005 was used and registered as significant (according to the presentation in tables of critical values). However, NOTE that this means a stretching of the p-value for significance up to 0.2 and that usually p-values above p=0.05 is not regarded as a good reliable significant result. No statistical calculations were performed on the further evaluations of poor performing products, i.e. the 9-point hedonic scales with added attributes and the Just-about-right-scales.

6.6. Screening of participants

Screening of participants was done before inclusion in the sensory panels using a questionnaire together with a practical test where participants were asked to identify the five basic tastes. Inclusion criteria were specified as:

- Liking the type of product
- Participant belonging to the target market
- Participant being a regular user of the product
- Participant sensory acuity (i.e. could identify five basic tastes; salt, sour, bitter, sweet and umami, and could identify a neutral sample)

Note that the questionnaire for the screening was developed and used by Unilever before the beginning of this project. However, further screening of employees has been conducted as part of this project, to increase the number of potential participants in the tests.

6.7. Invitations and instructions given to the participants prior to the tests

Invitations to the tests were sent out approximately one week before the test-session. The invitations included information regarding what to think about prior to the test i.e.

- Try not to smoke, drink coffee or eat spicy food too close to the session
- Try to avoid wearing strong smelling perfume, hairspray or similar
- Try to avoid being too full or too hungry at the tasting

A minimum of ten participating individuals per trial were decided. To get approximately ten participants at the test, invitations were sent out to at least the double number of employees. The participating employees were asked to arrive at the test anytime during a given hour.

6.8. Coding and order of samples

In the competitor tests the numbers, coding and order of the samples were performed according to standard sensory evaluation practice. A table of three-digit random numbers was used for the product coding. The codes were clearly placed on the samples without being very prominent. The order of the samples were to greatest extent balanced i.e. each sample

appeared in a given position an equal number of times. The pens/markers used for the coding were controlled not to deliver strong odours.

In the test for further evaluation of poor performing products only one single product was tested, why no number, coding or order needed to be performed.

6.9. Preparation and serving of products

The equipment and procedures used at the preparation of products and product presentation was carefully selected and monitored, following standard sensory evaluation practice. Efforts to reduce potential biases were made; the containers and equipment used was checked not to transferring any aroma or flavour to the product, and efforts to get all samples served at the same temperature and in equal amounts was done. All products tested were served together with suitable accompaniment, to mimic the manner they normally are consumed. The preparation procedures used was developed during the time of the project.

6.9.1. Preparation procedures used in the competitor tests

Table margarine

The margarines were kept in the fridge overnight prior to the test. The samples were coded, prepared and put in the fridge on trays a couple of hours before the test, to be ready and having the same temperature when participants arrived. To be sure all the samples had the same temperature when tasted, the trays with samples were taken out from the fridge where after participants arrived.

Since table margarines and butters often are consumed together with bread, bread was available at the tastings.

Baking margarine

To enable comparison of different brands of baking margarines, cookies/shortbreads were baked. The recipe was simple, with no added flavours, to allow the participants to taste the margarine in the best possible way. Since the participants were Finnish, the chosen recipe for the cookies was a classical Finnish recipe:

400g margarine
4 dl sugar
8 dl wheat flour
2tbs baking powder
2 eggs

The margarines were placed in room temperature a few hours before baking and the cookies were baked in exactly the same way; adding all ingredients together into a “kitchen aid”. When all ingredients were mixed, three pieces of three cm in diameter rolls were rolled, wrapped in plastic film and put into the fridge. After resting in the fridge the rolls were unwrapped and cut into approximately 1 cm pieces, and put on a baking tray in the oven at 180˚C for 6-8 minutes. The cookies were left to cool on a rack, and when cooled kept in tightly closing containers at room temperature.

Soup

The soups were prepared according to the instructions given on their packages. The brands on tube were heated in a pot just the way they were, and the canned brands were heated in a pot with added water. All pans were marked with the chosen code number, and to keep the same temperature on the samples the soup was poured into the prepared sample cups where after the participants arrived.

Depending on type of soup different suitable accompaniments was available for the participants.

Bouillon

One cube of bouillon from each brand was heated in pots together with 8dl water and 2dl cooking cream (15% fat). The reason for adding water and cream to the bouillon was to get the participants to feel the flavour in a more realistic way i.e. closer to the way they normally consume it, but without adding any other strong flavours. All pans were marked with the chosen code number, and to keep the same temperature on the samples the bouillons were poured into the prepared sample cups where after the participants arrived.

Sauce (dry mix)

The sauces were prepared according to the instructions given on their packages. Different brands had slightly different instructions, but all included adding a specified amount of milk and butter/margarine. The sauces were heated in pots on the stove. All pans were marked with the chosen code number, and to keep the same temperature on the samples the soups were poured into the prepared sample cups where after the participants arrived.

Tea

The tea tastings were prepared with special tea-testing containers/cups. These are specially developed for the preparation of tea. One tea bag was put in each container and 2dl water was added. The lid was put on and the tea left to stand for approximately 3 minutes (according to instructions on the tea package). The tea was then pored over to a bowl or a thermos, from which the tea was served directly into the coded sample cups and handed out to the participants where after they arrived. All containers used were carefully marked with the code number to ensure that no samples were mixed up.

Since tea often is consumed together with milk and/or sugar, the participants had milk and sugar as optional accompaniments available at the tastings.

6.9.2. Preparation procedures used in the test evaluating poor performing products

Soup 1

The company’s own brand of soup was prepared according to instructions given on the package. The soup was poured into the prepared sample cups where after the participants arrived.

6.10. Questionnaires; instructions and personal information

All questionnaires were given in English and constructed to be self-instructional to enable the participants to arrive at any time during a given hour. The questionnaires consisted of a front page that provided general information regarding what types of tests that were to be performed, what kind of information that was requested, and what to think about when performing the test. The participants were asked to answer the questions as a representative of the consuming population and to give their own, personal opinion regarding the food tested. They were instructed not to talk to each other during the test. Moreover, the front page also gave a short presentation of the type of product that was to be tested and what accompaniment that was available. Every test came with additional short and clear instructions. The last page in the questionnaire contained questions regarding the participating individual; gender, age group, frequency of consumption and appeal for the product. For more detailed facts regarding the layout and information given in the questionnaires, see appendix 1.

7. Results

7.1. Results from competitor tests

Below follow the results for each competitor test performed. The number of individuals participating in the tests is stated in connection with each test. Compressed results are presented in *Table 1, 2, 3 and 4,* found in the end of this section. An example of how the results from the evaluations were analyzed in detail is found in appendix 2.

7.1.1 Margarine 1 - Sweden

Evaluation of: Margarine A (Unilever), Margarine B and Margarine C

- Hedonic test (n=10): All brands had the majority of votes located in "the liking part" of the hedonic scale, and they all got a median value of 3 (= "like moderately"). The T-test showed no significant difference in liking between the brands *(Table 1 and Table 2).*

- Ranking test (n=10): The own brand (Margarine A) got scored as most preferred the largest number of times (*Figure 1*), and got the best median value. The Friedman test showed no significant difference in preference between the brands (*Table 1).*

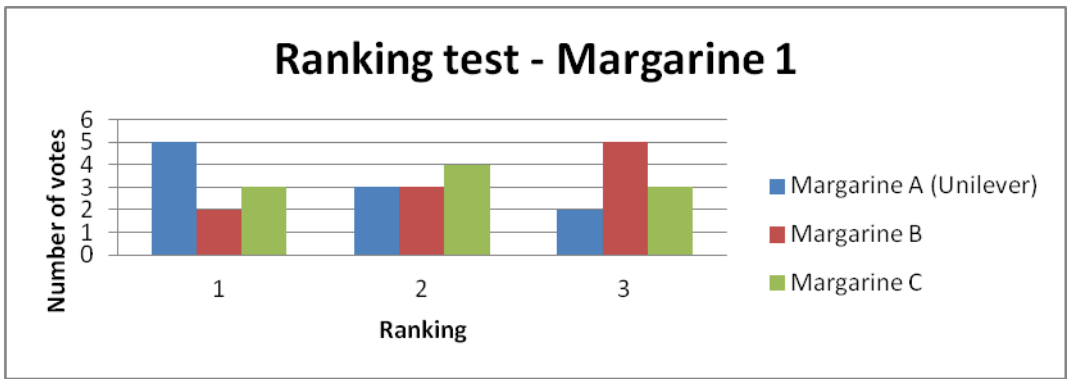

Figure 1. Results from the ranking test of margarine (where 1= most preferred and 3 = least preferred).

7.1.2. Margarine 2 - Sweden

Evaluation of: Margarine A (Unilever), Margarine B (Unilever), Margarine C, Margarine D and Margarine E

- Hedonic (n=10): Margarine A and E had the majority of votes located in "the liking part" of the hedonic scale, while the opinions regarding Margarine B, C and D differed among the participants. Margarine A got the best median value; 2 = "like very much", while Margarine B got the worst; 6 = "dislike slightly". The T-test showed that Margarine A was significantly more liked compared to Margarine C and D *(Table 1)*. The T-test also showed that Margarine B was significantly less liked compared to all the other brands *(Table 2)*.

- Ranking (n=10): Margarine A got the largest number of votes for most preferred, and Margarine B got the largest number of votes for least preferred *(Figure 2)*. Margarine A got the best median value, while Margarine B got the worst. The Friedman test showed that there was a significant difference in preference between the brands, and the LSD showed that Margarine B was significantly less preferred compared to all the other brands *(Table 2)*. However, no significance regarding if Margarine A was more preferred over the competitor brands was seen *(Table 1)*.

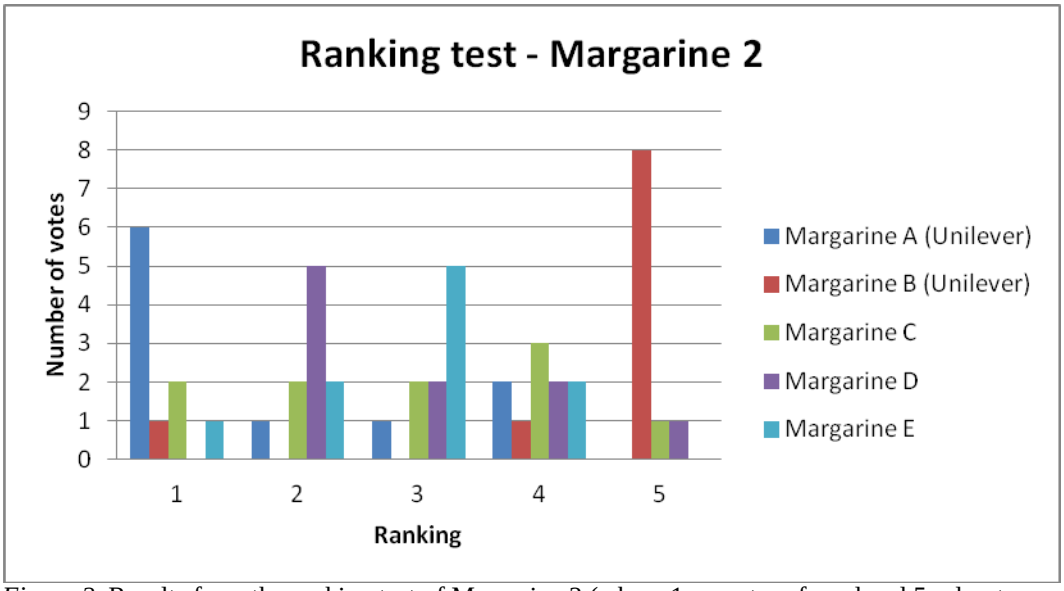

Figure 2. Results from the ranking test of Margarine 2 (where 1= most preferred and 5 = least preferred).

7.1.3. Margarine 3 - Sweden

Evaluation of: Margarine A (Unilever), Margarine B (Unilever), Margarine C, Margarine D and Margarine E

- Hedonic (n=23): The opinions regarding all the brands differed among the participants. For all brands, except for Margarine C, the majority of the votes were located in "the liking part" of the hedonic scale. However, Margarine C had the majority of votes located in "the disliking part" of the hedonic scale. Margarine A and E got the best median values; 3 = "like moderately", followed by Margarine B and D; 4 = "like slightly", and Margarine C; 6 = "dislike slightly". The T test showed that both of the own brands, Margarine A and B, were significantly more

liked compared to Margarine C. *(Table 1).* No other significant difference in liking between the own brands and the competitor brands was seen *(Table 1 and 2).*

- Ranking (n=21): The opinions regarding Margarine A, B and E differed among the participants, while Margarine C got the majority of votes located in "the less preferred part" of the diagram, and Margarine D got the majority of votes located in "the more preferred part" of the diagram *(Figure 3).* Margarine D and E got the best median values, followed by Margarine A, B and C. The Friedman test showed no significant difference in preference between the brands *(Table 1 and 2).*

  The calculations comprising all participants followed the results presented above i.e. the result from the calculations comprising only the participants that had verified they like table margarine *(Table 3 and 4).*

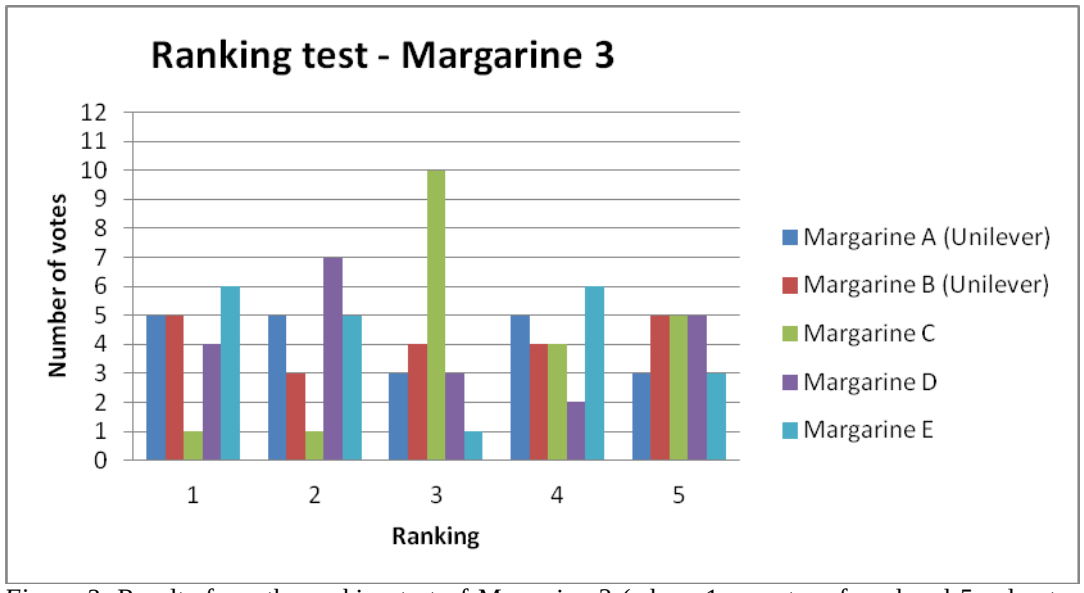

Figure 3. Results from the ranking test of Margarine 3 (where 1= most preferred and 5 = least preferred).

7.1.4. Margarine 4 – Finland

Evaluation of: Margarine A (Unilever), Margarine B, Margarine C and Margarine D

- Hedonic (n=17): The opinions regarding Margarine A and C differed among the participants, while the majority of votes for Margarine B and D were placed in "the liking part" of the hedonic scale. The median values for Margarine A and C were 6 = "dislike slightly", while Margarine B and D got 3 = "like moderately". The T-test showed that the own brand, Margarine A, was significantly less liked compared to the competitor brands Margarine B, and D *(Table 2).*

- Ranking (n=17): Margarine B got the largest number of most preferred votes, and Margarine A the largest number of least preferred votes *(Figure 4).* The Friedman test showed that there was a significant difference in preference between the brands, and LSD showed that the own brand, Margarine A, was significantly less preferred compared to Margarine B and D *(Table 2).*

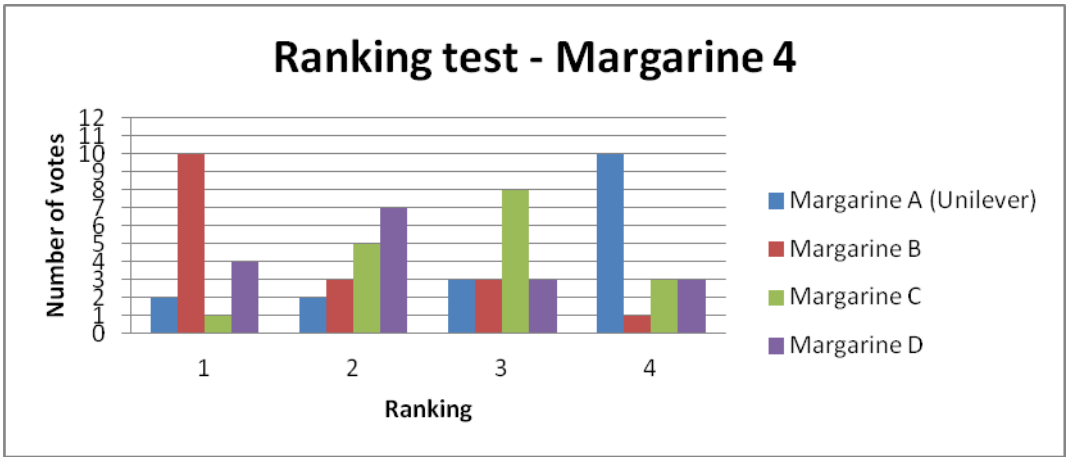

Figure 4. Results from ranking test of Margarine 4 (where 1 = most preferred and 4 = least preferred).

7.1.5. Margarine 5 - Finland

Evaluation of: Margarine A (Unilever), Margarine B (Unilever) and Margarine C

- Hedonic (n=25): All the brands got the majority of their votes in "the liking side" of the hedonic scale, and all brands got a median value of 3 (= "like moderately"). The T-test showed that the own brand, Margarine B, was significantly more liked over Margarine C *(Table 1).*No significant difference in liking between the own brand, Margarine A, and Margarine C was seen *(Table 1 and 2).*

- Ranking (n=25): The own brand, Margarine B, got the largest number of most preferred votes, and the own brand, Margarine A, got the largest number of least preferred votes *(Figure 5).* The Friedman test showed no significant difference in preference between the brands *(Table 1 and 2).*

  The calculations comprising all participants followed the results presented above, i.e. the results from the calculations comprising only the participants that had verified they liked margarine. However, the significance value shown in the hedonic test was stronger when comprising all participants compared to when only comprising the once liking *(Table 1 and 3).*

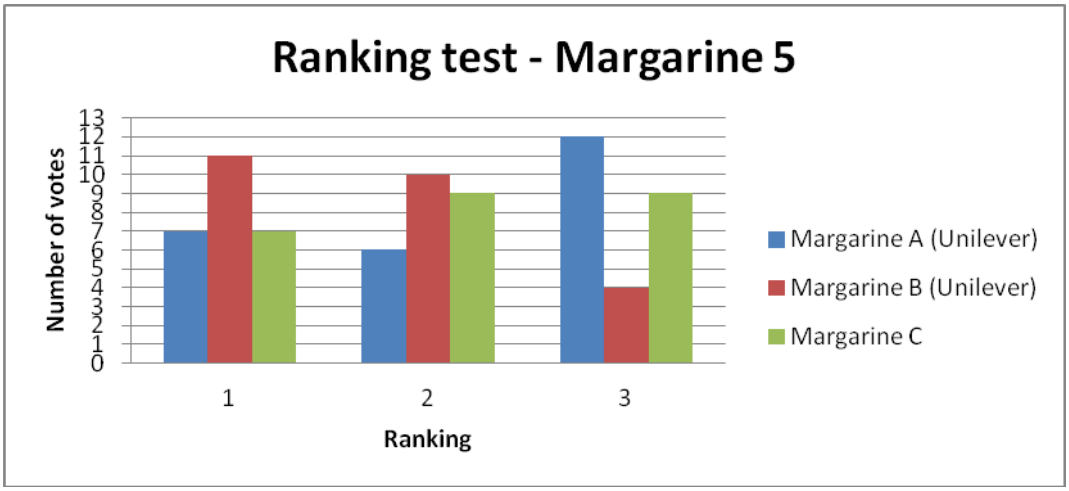

Figure 5. Results from ranking test of Margarine 5 (where 1= most preferred and 3= least preferred).

7.1.6. Soup 1 - Sweden

Evaluation of: Soup A (Unilever), Soup B, Soup C and Soup D

- Hedonic (n=10): Soup A and D got the majority of votes in "the disliking part" of the hedonic scale, while Soup B and C got the majority of votes in "the liking part" of the hedonic scale. Soup A and D got median values above five, i.e. in "the disliking side" of the hedonic scale. Soup B and C got median values of 3 (= "like moderately"). The T-test showed that Soup A was significantly less liked compared to Soup B, and C *(Table 2).* No significant difference in liking could be seen between the own brand, Soup A, and Soup D *(Table 2).*

- Ranking (n=10): Soup A and D got all, or almost all, votes in "the less preferred part" of the diagram, while Soup B and C got all, or almost all, votes in "the more preferred part" of the diagram *(Figure 6).* The Friedman test showed that there was a significant difference in preference between the brands, and LSD showed that the own brand, Soup A, was significantly less preferred compared to Soup B and C *(Table 2).* No Significant difference in preference between Soup A and D could be seen *(Table 2).*

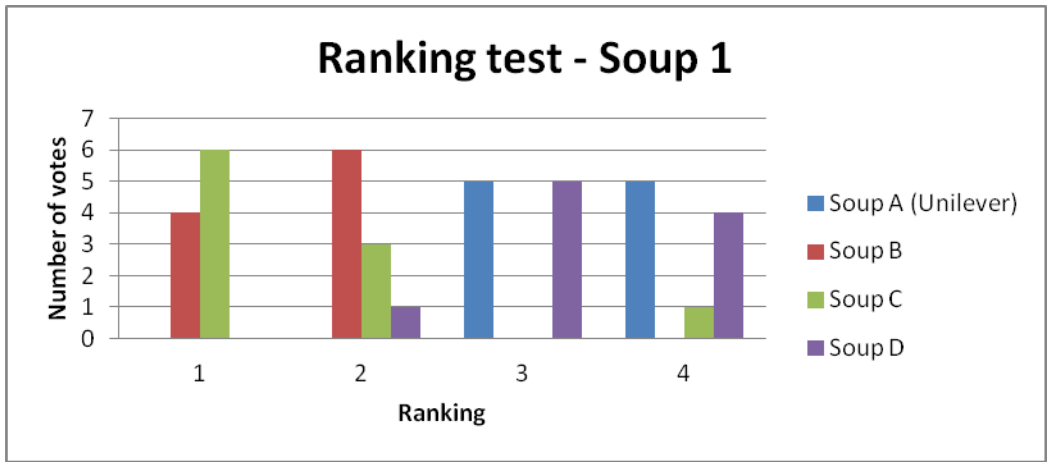

Figure 6. Results from ranking test of Soup 1 (where 1= most preferred and 4 = least preferred).

7.1.7. Soup 2 - Sweden

Evaluation of; Soup A (Unilever), Soup B and Soup C

- Hedonic (n=10): Soup A and C got the majority of votes in "the liking part" of the hedonic scale, while the opinions regarding Soup B differed among the participants. Soup A got the best median value (3="like moderately") followed by Soup B and C (4 = "like slightly"). The T-test showed that the own brand, Soup A, was significantly more liked compared to Soup B. No significant difference in liking could be seen between Soup A and C *(Table 1).*

- Ranking (n=10): The own brand, Soup A, got the largest number of most preferred votes, and the smallest number of least preferred votes *(Figure 7).* The Friedman test showed no significant difference in preference between the brands *(Table 1).*

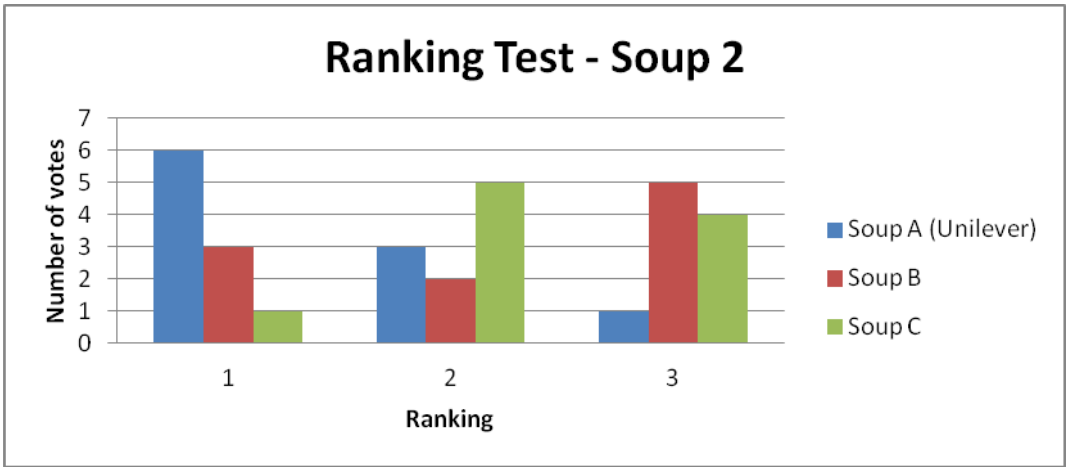

Figure 7. Results from ranking test of Soup 2 (where 1 = most preferred and 3 = least preferred).

7.1.8. Sauce 1 - Sweden

Evaluation of; Sauce A (Unilever), Sauce B and, Sauce C

- Hedonic (n=10): The opinions regarding Sauce A and B differed among the participants, while the majority of votes for Sauce C were placed in "the liking part" of the hedonic scale. Sauce A and B got median values above five, i.e. in "the disliking side" of the hedonic scale, while Sauce C got a median value of 2 (="like very much"). The T-test showed that the own brand, Sauce A, was significantly less liked compared to Sauce C *(Table 2).* No significant difference in liking between Sauce A and B could be seen *(Table 1 and 2).*

- Ranking (n=10): Sauce C got the largest number of most preferred votes and no votes for least preferred, while Sauce B got the largest number of least preferred votes *(Figure 8).* The Friedman test showed that there was a significant difference in preference between the brands, and the LSD showed that the own brand, Sauce A was significantly less preferred compared to Sauce C *(Table 2).* No significant difference in preference between Sauce A and B could be seen *(Table 1 and 2).*

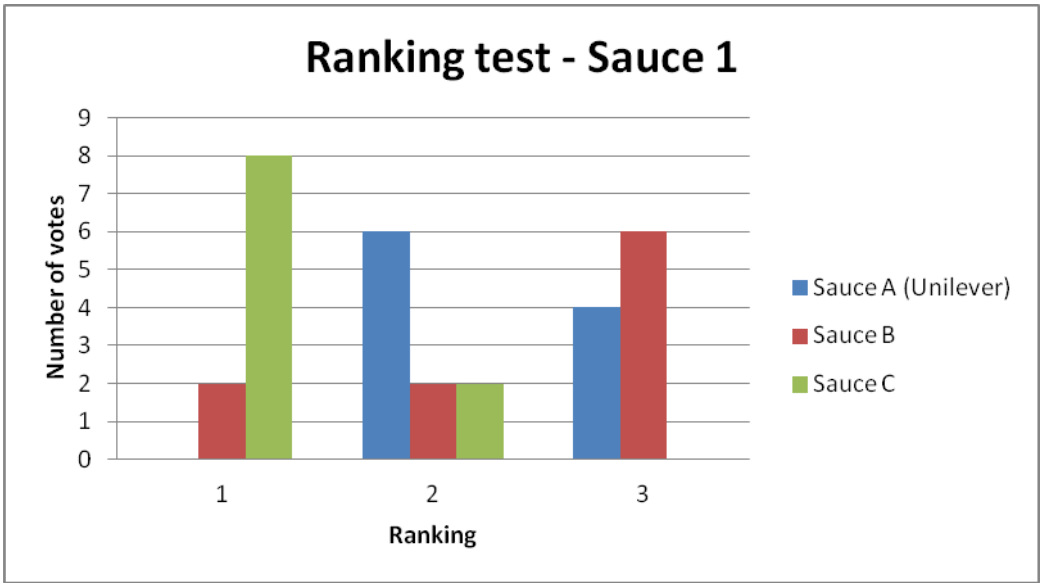

Figure 8. Results from ranking test of Sauce 1 (where 1= most preferred and 3 = least preferred).

7.1.9. Bouillon 1 - Finland

Evaluation of: Bouillon A (Unilever) and Bouillon B

- Hedonic (n=10): The own brand, Bouillon A, got the majority of votes in "the liking part" of the hedonic scale, while the opinions regarding Bouillon B differed among the participants. Bouillon A got the best median value (3 = "like moderately") compared to Bouillon B (4 ="like slightly"). The T-test showed no significant difference in liking between Bouillon A and B *(Table 1).*

- Ranking (n=10): Six out of ten participants ranked the own brand, Bouillon A, as most preferred over Bouillon B *(Figure 9).* The Friedman test showed no significant difference in preference between Bouillon A and B *(Table 1).*

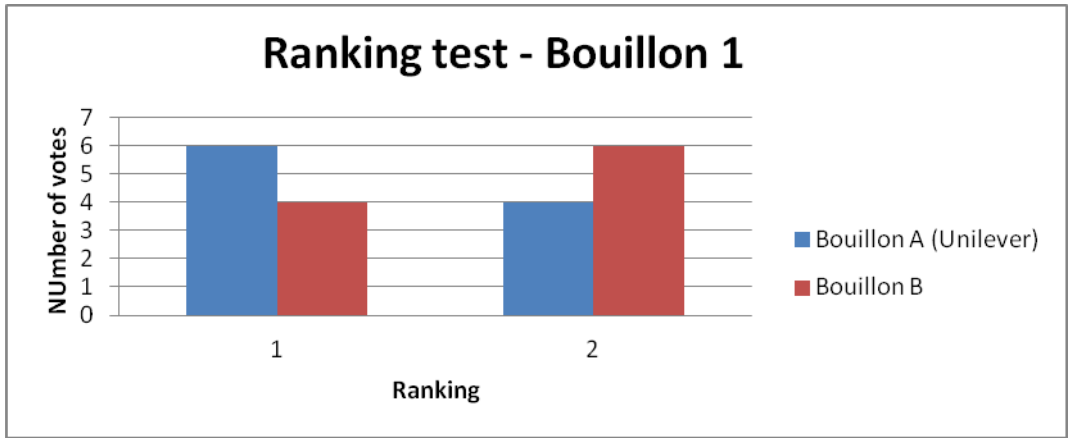

Figure 9. Results from ranking test of Bouillon 1 (where 1= most preferred and 2 = least preferred).

7.1.10. Tea 1 - Sweden

Evaluation of: Tea A (Unilever), Tea B (Unilever), Tea C (Unilever), Tea D and Tea E

- Hedonic (n=12): The own brands, Tea A, B and C, all got the majority of votes in "the liking part" of the hedonic scale, while the opinions regarding Tea D and E differed among the participants. Tea A, B and C, got the best mean and median values, followed by Tea D, and last E. All brands got median values in "the liking side" of the hedonic scale. The T-test showed that all the own brands, Tea A, B and C, were significantly more liked compared to Tea E *(Table 1).* The own brands, Tea A and C, also showed to be significantly more liked compared to Tea D *(Table 1).*

- Ranking (n=11): The opinions regarding the own brands, Tea A, B and C, differed among the participants. Tea D got the majority of the votes in "the more preferred part" of the diagram, and Tea E got the majority of votes in "the less preferred part" of the diagram *(Figure 10).* All brands, except for Tea E, got similar mean and median values. The mean and median value of Tea E was slightly higher compared to the others. The Friedman test showed no significant difference in preference between the brands *(Table 1 and 2).*

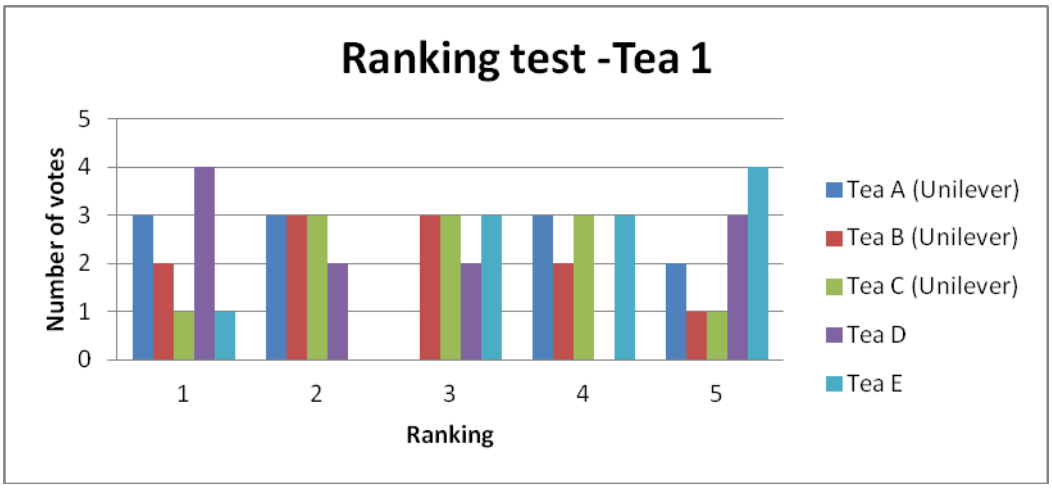

Figure 10. Results from ranking test of Tea 1 (where 1= most preferred and 5 = least preferred).

7.1.11. Tea 2 - Sweden

Evaluation of: Tea A (Unilever), Tea B (Unilever), Tea C, Tea D and Tea E

- Hedonic (n=12): Tea A, B and C all had the majority of votes in "the liking part" of the hedonic scale. Tea D and E had the majority of the votes in "the disliking part" of the hedonic scale. Tea A, B and C all got median values in "the liking side" of the hedonic scale, Tea D and E got median values in "the disliking side" of the hedonic scale. The T-test showed that both own brands, Tea A and B, were significantly more liked compared to Tea D and E *(Table 1).* No significant difference in liking between the own brands and Tea C was seen *(Table 1 and 2).*

- Ranking (n=12): Tea A, B and C all had the majority of votes in "the more preferred" part of the diagram, while Tea D and E got the majority of votes in "the less preferred part" of the diagram. Tea D got eight votes for least preferred *(Figure 11).* The Friedman test showed that Tea A and B were significantly more preferred over Tea D and E. No significant difference in liking between the own brands and Tea C was seen *(Table 1).*

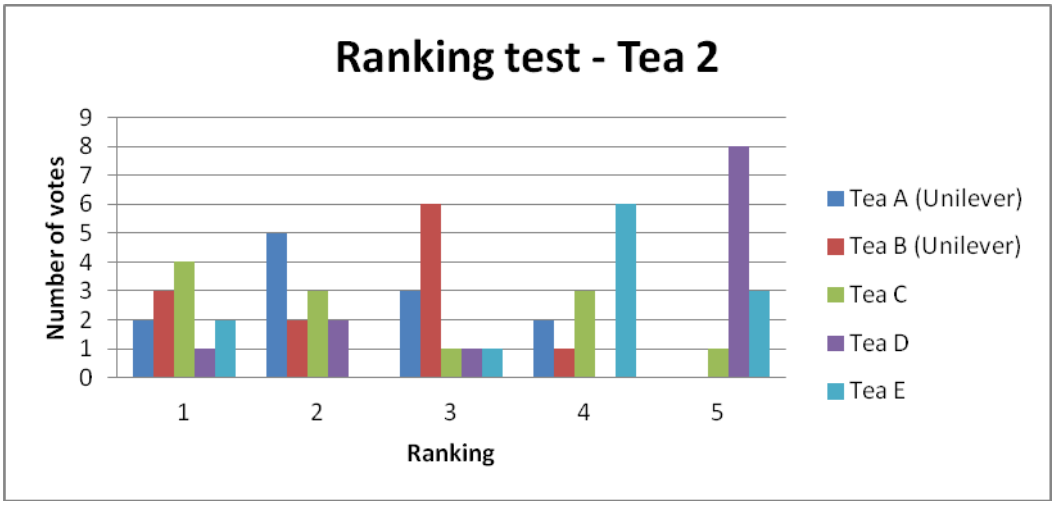

Figure 11. Results from ranking test of Tea 2 (where 1= most preferred and 5 = least preferred).

7.1.12. Tea 3 - Sweden

Evaluation of: Tea A (Unilever) and Tea B

- Hedonic (n=14): Tea A and B both had the majority of votes in "the liking part" of the hedonic scale, and both brands got a median value below 5, i.e. in "the liking part" of the hedonic scale. The T-test showed no significant difference in liking between Tea A and B. *(Table 1)*

- Ranking: No ranking test performed.

  The calculations comprising all participants followed the results presented above, i.e. the results from the calculations comprising only the participants that had verified they like the type of tea. However a significant difference in liking was seen in the calculations comprising all participants. Tea A was significantly more liked over Tea B *(Table 3).*

7.1.13. Tea 4 - Finland

Evaluation of; Tea A (Unilever) and Tea B

- Hedonic (n=8): The majority of votes for the own brand, Tea A, was located in "the liking part" of the hedonic scale, while the opinions regarding Tea B differed among the participants. Tea A got a slightly better median value compared to Tea B. However, median values for both brands were located in "the liking part" of the hedonic scale. The T-test showed no significant difference in liking between Tea A and B *(Table 1).*

- Ranking (n=9): Six out of nine participants ranked Tea A as most preferred over Tea B *(Figure 12).* The Friedman test showed no significant difference in preference between Tea A and B *(Table 1).*

  The calculations comprising all participants followed the results presented above, i.e. the results from the calculations comprising only the participants that had verified they like the type of tea tested *(Table 1 and 3).*

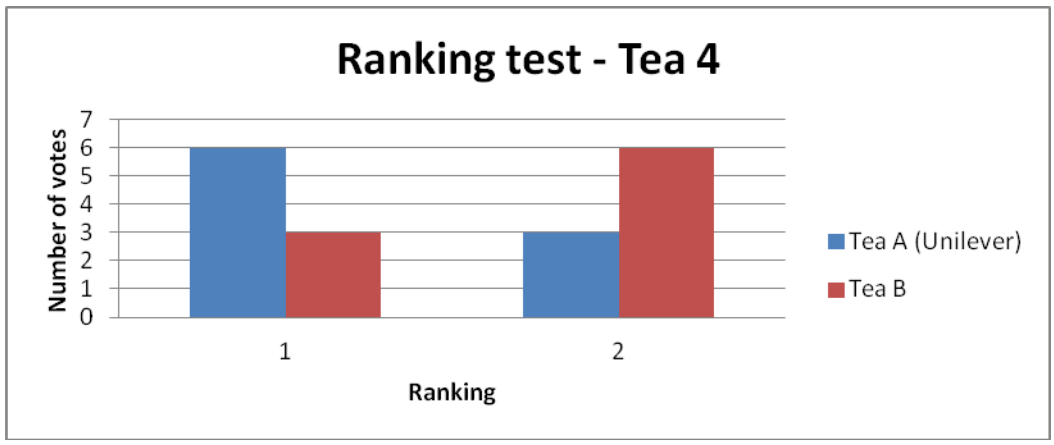

Figure 12. Results from ranking test of Tea 4 (where 1= most preferred and 2 = least preferred).

7.1.14. Tea 5 - Finland

Evaluation of: Tea A (Unilever) and Tea B

- Hedonic (n=12): The opinions regarding liking of both brands differed among participants. The mean and median values were similar, and both brands had median values placed in "the liking side" of the diagram. The T-test showed no significant difference in liking between the brands. *(Table 1)*

- Ranking (n=13): Tea A and B got almost the same number of most and least preferred votes *(Figure 13).* The Friedman test showed no significant difference in preference between the brands *(Table 1).*

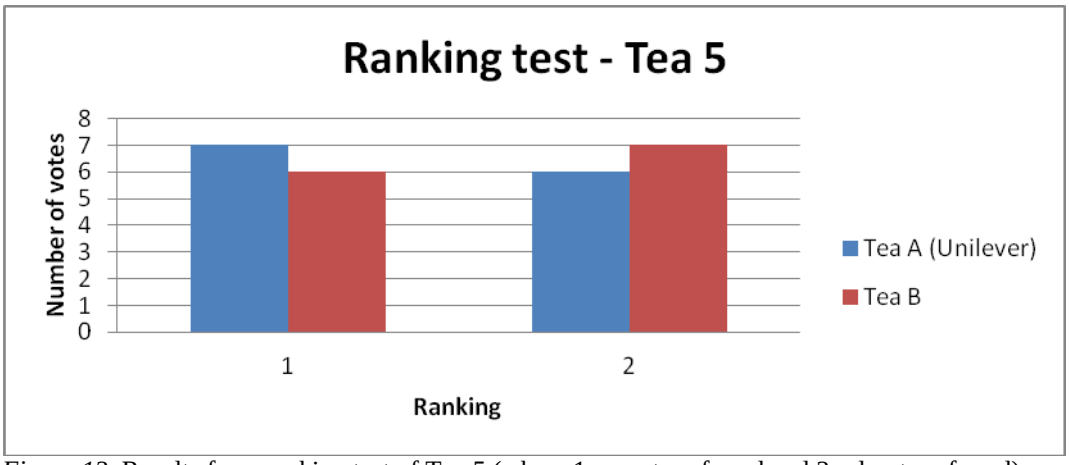

Figure 13. Results from ranking test of Tea 5 (where 1= most preferred and 2 = least preferred).

7.1.15. Tea 6 - Finland

Evaluation of: Tea A (Unilever), Tea B (Unilever) and Tea C

- Hedonic (n=11): The opinions regarding all of the brands differed among the participants. However, Tea A got the majority of votes in "the liking part" of the hedonic scale. Tea A got the best median value (3= "like moderately"), followed by Tea C (5 = "neither like nor dislike") and last Tea B (6 = "dislike slightly"). The T-test showed that the own brand, Tea A, was significantly more liked over Tea C *(table 1).* No significant difference in liking between Tea B and C could be seen *(Table 2).*

- Ranking (n=10): The own brand, Tea A, got no votes for least preferred, while the own brand, Tea B, got the largest number of least preferred votes and the smallest number of most preferred votes. The opinions regarding Tea C differed among the participants *(Figure 14).* The Friedman test showed no significant difference in preference between the brands *(Table 1 and 2).*

  The calculations comprising all participants followed the results presented above, i.e. the results from the calculations comprising only the participants that had verified they liked the tea of the test. However, the significance value shown in the hedonic test was stronger when comprising all participants compared to when only comprising the once liking *(Table 1 and 3).*

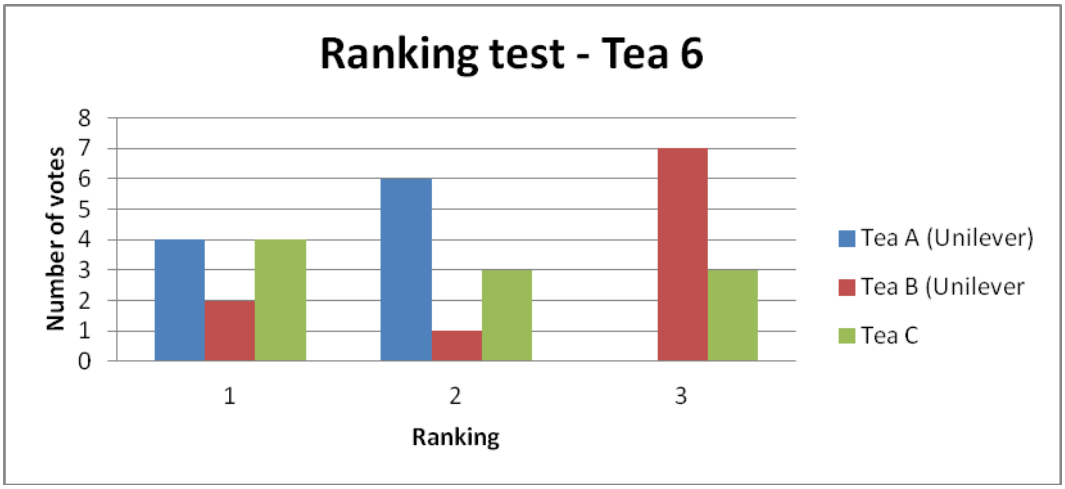

Figure 14. Results from ranking test of Tea 6 (where 1= most preferred and 3 = least preferred).

7.1.16. Tea 7 - Finland

Evaluation of: Tea A (Unilever) and Tea B

- Hedonic (n=10): Both brands have the majority of votes in "the liking part" of the hedonic scale, and got median values below five (Tea A; 4= "like slightly" and Tea B; 3 = "like moderately"). The T-test showed no significant difference in liking between the brands *(Table 2).*

- Ranking (n=10): Seven out of ten participants ranked Tea B as most preferred over the own brand, Tea A. *(Figure 15).* The Friedman Test showed no significant difference in preference between the samples *(Table 2).*

  The calculations comprising all participants followed the results presented above, i.e. the results from the calculations comprising only the participants that had verified they liked the type of tea of the test *(Table 2 and 4).*

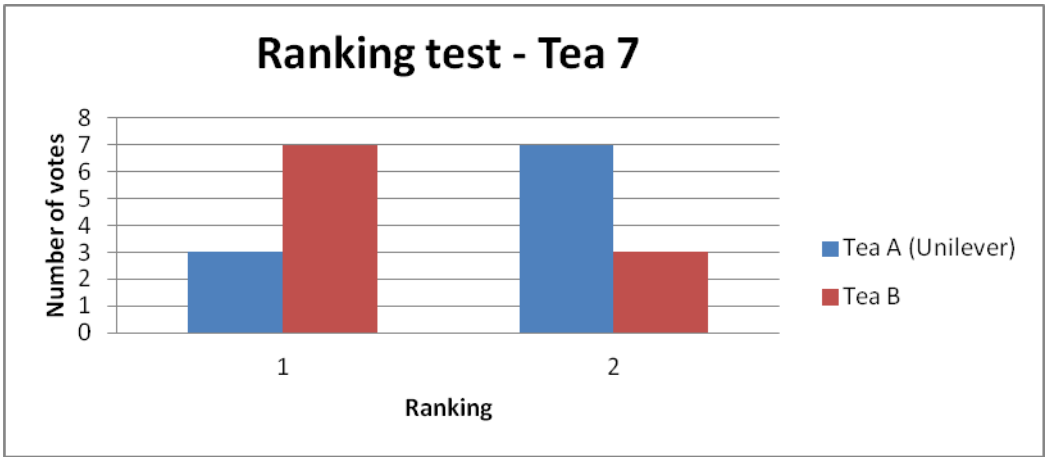

Figure 15. Results from ranking test of Tea 7 (where 1= most preferred and 2 = least preferred).

Table 1. Summarized results from the 9-point hedonic scale test and the ranking test comprising participants verifying they like the food tested. The table shows Unilever brands potentially significantly more liked/preferred compared to competitor brands (OB =Own brand, CB = Competitor brand, NS = No significance)

			Hedoic test				Ranking test			
Evaluation	Own Brand	Competitor brand	n	Median	Median	T-test	n	Median	Median	Friedman test
				OB	CB			OB	CB	
Margarine 1	Margarine A	Margarine B	10	3	3	NS	10	1,5	2,5	NS
	Margarine A	Margarine C	10	3	3	NS	10	1,5	2	NS
Margarine 2	Margarine A	Margarine C	10	2	5	0,2	10	1	3	NS
	Margarine A	Margarin D	10	2	3,5	0,05	10	1	2,5	NS
	Margarine A	Margarine E	10	2	3,5	NS	10	1	3	NS
Margarine 3	Margarine A	Margarine C	23	3	6	0,02	21	3	3	NS
	Margarine A	Margarine D	23	3	4	NS	21	3	2	NS
	Margarine A	Margarine E	23	3	3	NS	21	3	2	NS
	Margarine B	Margarine C	23	4	6	0,1	21	3	3	NS
Margarine 5	Margarine A	Margarine C	25	3	3	NS	25	2	2	NS
	Margarien B	Margarine C	25	3	3	0,1	25	2	2	NS
Soup 2	Soup A	Soup B	10	3	4	0,2	10	1	2,5	NS
	Soup A	Soup C	10	3	4	NS	10	1	2	NS
Sauce 1	Sauce A	Sauce B	10	6	5,5	NS	10	2	3	NS
Bouillon 1	Bouillon A	Bouillon B	10	3	4	NS	10	1	2	NS
Tea 1	Tea A	Tea D	12	4	4	0,2	11	2	2	NS
	Tea A	Tea E	12	4	5	0,05	11	2	4	NS
	Tea B	Tea D	12	3,5	4	NS	11	3	2	NS
	Tea B	Tea E	12	3,5	5	0,1	11	3	4	NS
	Tea C	Tea D	12	3,5	4	0,2	11	3	2	NS
	Tea C	Tea E	12	3,5	4	0,1	11	3	4	NS
Tea 2	Tea A	Tea C	12	3,5	4	NS	12	2	2	NS
	Tea A	Tea D	12	3,5	7,5	0,002	12	2	5	0,025
	Tea A	Tea E	12	3,5	5,5	0,05	12	2	4	0,025
	Tea B	Tea C	12	4,5	4	NS	12	3	2	NS
	Tea B	Tea D	12	4,5	7,5	0,05	12	3	5	0,025
	Tea B	Tea E	12	4,5	5,5	0,1	12	3	4	0,025
Tea 3	Tea A	Tea B	14	2,5	3,5	NS	x	x	x	x
Tea 4	Tea A	Tea B	8	3,5	4,5	NS	9	1	2	NS
Tea 5	Tea A	Tea B	12	4	4,5	NS	13	1	2	NS
Tea 6	Tea A	Tea C	11	3	5	0,2	10	2	2	NS

Table 2. Summarized results from the 9-point hedonic scale test and the ranking test comprising participants verifying they like the food tested. The table shows Unilever brands

possibly significantly less liked/preferred compared to competitor brands (OB = Own brand, CB = Competitor brand, NS = No significance).

			Hedonic test				Ranking test			
Evaluation	Own Brand	ompetitor bran	n	Median	Median	T-test	n	Median	Median	Friedman test
				OB	CB			OB	CB	
Margarine 1	Margarine A	Margarine C	10	3	3	NS	10	1,5	2	NS
Margarine 2	Margarine B	Margarine C	10	6	5	0,1	10	5	3	0,01
	Margarine B	Margarin D	10	6	3,5	0,1	10	5	2,5	0,01
	Margarine B	Margarine E	10	6	3,5	0,01	10	5	3	0,01
Margarine 3	Margarine A	Margarine E	23	3	3	NS	21	3	2	NS
	Margarine B	Margarine D	23	4	4	NS	21	3	2	NS
	Margarine B	Margarien E	23	4	3	NS	21	3	2	NS
Margarine 4	Margarine A	Margarine B	17	6	3	0,01	17	4	1	0,005
	Margarine A	Margarine C	17	6	6	NS	17	4	3	NS
	Margarine A	Margarine D	17	6	3	0,05	17	4	2	0,005
Margarine 5	Margarine A	Margarine C	25	3	3	NS	25	2	2	NS
Soup 1	Soup A	Soup B	10	6,5	3	0,002	10	3,5	2	0,05
	Soup A	Soup C	10	6,5	3	0,001	10	3,5	1	0,05
	Soup A	Soup D	10	6,5	6	NS	10	3,5	3	NS
Sauce 1	Sauce A	Sauce B	10	6	5,5	NS	10	2	3	NS
	Sauce A	Sauce C	10	6	2	0,01	10	2	1	0,01
Tea 1	Tea A	Tea D	12	4	4	NS	11	2	2	NS
	Tea B	Tea D	12	3,5	4	NS	11	3	2	NS
	Tea C	Tea D	12	3,5	4	NS	11	3	2	NS
Tea 2	Tea B	Tea C	12	4,5	4	NS	12	3	2	NS
Tea 6	Tea B	Tea C	11	6	5	NS	10	3	2	NS
Tea 7	Tea A	Tea B	10	4	3	NS	10	2	1	NS

Table 3. Summarized results from the 9-point hedonic test and the ranking test comprising all participants. The table shows Unilever brands potentially significantly more liked/preferred compared to competitor brands (OB = Own brand, CB = Competitor brand, NS = No significance).

			Hedoic test				Ranking test			
Evaluation	Own Brand	Competitor brand	n	Median	Median	T-test	n	Median	Median	Friedman test
				OB	CB			OB	CB	
Margarine 3	Margarine A	Margarine C	30	3,5	5	0,02	26	3	3	NS
	Margarine A	Margarine D	30	3,5	4	NS	26	3	2	NS
	Margarine B	Margarien C	30	4	5	0,05	26	3	3	NS
	Margarine B	Margarine D	30	4	4	NS	26	3	2	NS
Margarine 5	Margarine A	Margarine C	31	4	3	NS	30	2	2	NS
	Margarine B	Margarine C	31	3	3	0,05	30	2	2	NS
Tea 3	Tea A	Tea B	27	3	4	0,2	x	x	x	x
Tea 4	Tea A	Tea B	12	3,5	3,5	NS	13	1	2	NS
Tea 6	Tea A	Tea C	13	4	5	0,1	12	2	2	NS
Tea 7	Tea A	Tea B	13	4	3	NS	12	2	1	NS

Table 4. Summarized results from the 9-point hedonic test and the ranking test comprising all participants. The table shows Unilever brands possibly significantly less liked/preferred

compared to competitor brands (OB = Own brand, CB = Competitor brand, NS = No significance).

			Hedoic test				Ranking test			
Evaluation	Own Brand	Competitor brand	n	Median	Median	T-test	n	Median	Median	Friedman test
				OB	CB			OB	CB	
Margarine 3	Margarine A	Margarine D	30	3,5	4	NS	26	3	2	NS
	Margarine A	Margarine E	30	3,5	3	NS	26	3	2,5	NS
	Margarine B	Margarine D	30	4	4	NS	26	3	2	NS
	Margarine B	Margarine E	30	4	3	NS	26	3	2,5	NS
Margarine 5	Margarine A	Margarine C	31	4	3	NS	30	2	2	NS
Tea 6	Tea B	Tea C	13	6	5	NS	30	3	2	NS
Tea 7	Tea A	Tea B	13	4	3	NS	12	2	1	NS

7.2. Results from further evaluations of a poor performing product – Soup 1

Hedonic scales: All attributes, except for “smell/aroma” got the majority of votes in “the liking part” of the hedonic scale and a median value of 3 = “like moderately”. Smell/aroma however got the majority of votes in the “disliking side” of the hedonic scale and a median value of 6 = “dislike slightly”.

JAR-scales: All attributes, except for “amount of herbs” and size of meat pieces, got the majority of votes on “just about right”. “Amount of herbs” however got the majority of votes on “not enough herbs”, and “size of meat pieces” got the majority of votes on “slightly too small”.

For more information regarding the results from the further evaluation of Soup 1A see appendix 3.

8. Discussion

8.1. Method and practices

The practical prerequisites for this project were:

- Use employees at the company as participants in the tests
- Have a number of approximately ten participating employees in each test
- Enable the participating employees to arrive at the test at any time during a given hour (drop-in test)
- Be able to use the test methods and questionnaires developed on different food categories without any larger modifications.
- Use and adjust the available facilities to serve as testing and preparation area

8.1.2. Using employees as participants

Using employees as participants includes both Pros and Cons (Meilgaard et al., 2007). The employees participating in the tests all worked at the food compartment of the company and thereby had good knowledge about the Unilever food products. Some of them had participated in a number of descriptive sensory evaluation tests before, why they were not perfectly qualified for representing average consumers. Efforts were made to remind the participating employees to think from a consumer point of view, to put their professional self aside and focus on their own personal opinion in preference and liking. Despite this, in the interpreting of results it showed that some participants used a more analytical way of evaluating the food compared to the average consumer. Participants used to perform descriptive tests on food products tended to in their comments forget to tell their affective opinion, and instead describe the sensory attributes. By if in the future getting employees from other compartments at the company, like homecare and hygiene, to participate in the sensory tests, negative biases could be limited. Another way to ensure that participating employees are qualified to represent the consuming population is to compare the internal consumer panel with an outside sample of non-employee consumers by testing the same product in both groups. This is a common advice when choosing to use internal consumer tests (Lawless and Heymann, 2010). However, continuous usage of the same employees as participants in the tests could possibly over time make the participants more analytical in their way of evaluating compared to the average consumer. Another important factor to consider is that using employees as participants in a consumer study makes it impossible to sample properly from the consuming population (Meilgaard et al., 2007). In this case the majority of employees participating in the tests were women aged 25-40.

8.1.3. The number of participants in the tests

Since consumer test often comprises a large number of participants, the suitability of using only approximately ten participants in the consumer tests was one of the big questions when starting the project. However, tendencies in liking and preference between the products have been seen despite the low number of participants, and significant differences in liking and preference was observed. The screening of participant's sensory acuity made sure that the small number of participants used all had the same sensory perception that the majority of consumers possess. This increased the chance of getting accurate and reliable results, as did the documenting of the participating employee's food consumption habits. However during the time of conducting the tests it became obvious that the manuals used must be more specified in the future, e.g. liking tea, does not necessary mean liking both fruit tea, black breakfast tea and peppermint tea. In some of the tests almost half of the participants showed not to like the sort of tea of the tasting, and therefore were not qualifying for the test. Despite this, in some cases calculations were made with all participants, since using only the number of individuals liking the tea tasted was not enough to draw any conclusions.

8.1.4. Drop-in test

The questionnaires used for the competitor tastings were constructed to be self instructive, to enable the participating employees to arrive at the test anytime during a given hour (a drop-in test). This proved to be a very time effective way to perform the tests, and it also contributed to the employee's flexibility. However, this design of the tests worked better for some food categories than others, e.g. for foods served and consumed cold, like margarines, when having an available fridge where prepared samples could be stored before serving. Hot foods,

like tea, that are easy to keep warm, for example by the use of a thermos also worked well. For other heated foods, like soups and sauces, a few difficulties arose. If kept warm for too long the consistency changed because water evaporated. This gives a risk that depending on what time participants arrived the texture of the soup/sauce could differ. Having a drop-in test during an hour may therefore be more or less suitable depending on the food product tested. Since the preparation procedures were developed during the time of the project, some of the evaluations performed did not follow the finally chosen preparation procedures, described under method. At the first two margarine evaluations: Margarine 1 and Margarine 2, no refrigerator was available, why the temperature and consistency of the margarine changed during the test. At two of the evaluations of tea: Tea 1 and Tea 2, no specially designed tea cups were available why the tea was prepared in pans on the stove.

8.1.5. Usage of methods and questionnaires

Unilever had a request that the test methods and questionnaires developed should be able to be used on different food categories without larger modifications. This request was met in the competitor questionnaires, however when performing more of a descriptive test to evaluate the reasons behind a bad result and what attributes that could be improved the questionnaires must be specially constructed to suit the product of interest. Nevertheless, it is possible to ease the work of designing these questionnaires by constructing lists of possible relevant attributes for each food category (See example in appendix 4)

8.1.6. Testing and preparation area

According to standard sensory evaluation practice regulated testing and preparation areas for the tests are important (Meilgaard et al., 2007). In lack of a custom-designed test area, the café at the company was used. The positive aspects with the usage of the café were that the location were in the centre of the working place and thereby easy to reach for all participating employees. However, the location had negative aspects in terms of; too much noise, people passing by, possible food odours etc. that could disturb the participants in their evaluation work. To limit these biases the time of the tests was carefully chosen. It appeared that the best times for conducting tests was in the morning around 9-10 or in the afternoon around 14-15. Depending on the type of food tested morning or afternoon could be chosen. In the lack of testing booths, small single tables were used, and participants asked not to communicate during the tests. To limit biases even more, improvements in the testing area could be made, or investments in a real testing area be done.

8.1.7. Reliability of the results

Performance of internal small scale consumer test requires adjustments and compromises in the sensory practices, why consideration regarding the reliability of the results is important.
It is hard to say to what extent the results are affected by the slightly compromised conditions at the company. To find out, a comparison of the results from the small scale internal consumer tests conducted at the company with the results from an outside test with real consumers as participants could be conducted. Comparison of results from a larger scale test, comprising a minimum of 50 participants, with the results from the same evaluation in small scale, approximately ten participants, at the company could also be a good way to make sure the small number of participants used is reliable. Nevertheless, since the purpose of the results of the small scale internal tests are not to be used as a proof, but only as a tool to help identify products that for some reason might be in need for further evaluations, the need for

precise evaluation conditions and practices are not as strictly essential. It will be up to the company to decide to what extent they believe that the results are reliable.

8.2. The 9-point hedonic scale and the ranking test - Statistics

The 9-point hedonic scale was chosen as one of the test methods because; there are large quantities of available experience due to its wide utilization (Tuorila, 2008) (Lawless and Heymann, 2010), it's been proven easy to use and understand for untrained participants, and it is relatively easy to compile the results (Lawless and Heymann, 2010). It also successfully complemented the ranking test, since the ranking test only tells what product that is preferred and not gives answers to the magnitude of preference (Meilgaard et al., 2007). An alternative to the ranking test could have been performance of Best-Worst-Scaling.

In the choice of statistical methods for calculation of results, a combination of advices from statisticians working at SLU and facts regarding common practices gathered from the literature was used. The usage of parametric statistics, t-test, when analyzing data from the 9-point hedonic scale has been debated in the literature (Lawless and Heymann, 2010). Even if the 9-point hedonic scale is specially constructed to achieve a true interval level of measurement, not all statisticians agree that the data gathered is true parametric. One common way of justifying the parametric approach is through the use of a larger sample size (Lawless and Heymann, 2010). In this project we can not justify the use of parametric statistics through a large group of participants. We will have to rely on the scale options obtaining sufficient equal differences. The primary reason to use the 9-point hedonic scale in this project was not to see if there was a significant difference in liking between brands, but to see the level of liking/disliking of each product. However, when analyzing the results from the ranking test calculations of significance was an important factor. Data gathered from ranking tests is ordinal and always treated as nonparametric (Lawless and Heymann, 2010). In the interpretation of the results it is therefore important to be aware of that the calculations deriving from the ranking test might be more reliable compared to the calculations deriving from the hedonic test. It is also important to reflect on that liking and preference can not be equated. Liking one product the most must not mean that it is most preferred (Meilgaard et al., 2007). e.g. you might like the sweetest dessert the most, but since you are on a diet you prefer the less sweet one. However, in many cases liking and preference do agree (Meilgaard et al., 2007).

For six of the competitor tests double calculations were made: one including only those participants that had verified they like the product tested, and one including all participants. For all the tests, the results from both calculations agreed (see at Results 7.1.3., 7.1.5., 7.1.12., 7.1.13., 7.1.15. and 7.1.16.). As expected in the calculations including a larger panel size i.e. in the calculations including all participants, it was proved to be easier to reach significant differences between the brands. This was seen in three out of the six calculations: Margarine 5, Tea 3 and Tea 6. One theory regarding why not using participants not liking the product tested was that they would influence the hedonic result by lowering the hedonic grade. To investigate if this is true mean values from the hedonic tests were compared. However, indications toward this being true were only shown in one out of six tests: Tea 3. The explanation to this could either be that liking is not that an important factor i.e. people not liking a product have the ability to grade it in a similar way to those liking it, or it could be due to that in addition to the participants disliking the product also participants that had not filled in the personal data of the questionnaires ended up in the calculations comprising all participants.

In the competitor tests, ten out of sixteen hedonic evaluations showed significant differences in liking between the brands. Seven of these sixteen evaluations included only ten participants. Among these seven, four evaluations showed significant differences in liking between the brands. Five out of fifteen ranking evaluations showed significant differences in preference between the brands. Among these five, three evaluations comprised only ten participants. This indicates that it is possible to perform both hedonic consumer tests and ranking tests and get significant results from panel groups comprising only ten participants. When comparing the results from the hedonic test and the ranking test, getting a significant difference between products appeared to be easier in the calculations using a t-test from the 9-point hedonic scale compared to using the Friedman test in combination with calculations of the Least Significant Difference from the ranking data.

8.3. Results from competitor tests – suggestions on how to proceed

Margarine 1 - Sweden

No significant difference in preference/liking was seen. However, the ranking test indicated a weak preference for the own brand why a larger scale ranking test could possibly give a significant "best in test" result.

Margarine 2 - Sweden

The own brand, Margarine B, was significantly less liked/preferred compared to all the competitor brands. This, in addition to a median value in the "disliking side" of the diagram makes Margarine B qualified for further tests too find the reason to the poor result. Comments from participants indicated that Margarine B leaves a film in the mouth and has not got a good taste.

The tests either indicated or showed that the own brand, Margarine A, was significantly more liked/preferred over all the other brands. A larger scale ranking test might show if Margarine A could be "best in test".

Margarine 3 – Sweden

All Unilever brands got median values in "the liking side of the diagram" and no indications of being less liked/preferred compared to the other brands were seen why no further evaluations are needed.

Margarine 4 – Finland

The tests showed that the own brand, Margarine A, was significantly less liked/preferred compared to two out of three competitor products. This, in addition to a median value of 6 ="dislike slightly" makes Margarine A qualify for further evaluations to find out the reason behind the bad result. Comments from the participants indicated that there is a problem with taste and texture.

Margarine 5 – Finland

Both own brands had median values in "the liking side" of the hedonic scale, and were either similar or significantly more liked compared to the competitor brand. No further evaluations are therefore necessary.

Soup 1 - Sweden
The own brand, Soup A, was significantly less liked/preferred compared to two out of three competitor brands. This, in addition to a median value of 6,5 =in between “dislike slightly” and “dislike moderately” makes Soup A qualify for further evaluations to find the reason behind the bad results. Comments from the participants indicated a problem regarding taste and consistency.

Soup 2 - Sweden
The results indicates that the own brand, Soup A, could be more liked/preferred compared to the competitor brands. A larger scale ranking test might show if Soup A could be “best in test”.

Sauce 1- Sweden
The own brand, Sauce A, was significantly less liked/preferred compared to one out of two competitor brands. This, in addition to a median value of 6 = “dislike slightly” makes it qualify for further tests to evaluate the reason behind the bad result. Comments from the participants indicated problems with taste and texture.

Bouillon - Finland
No significant difference in liking/preference was shown. However a weak indication of the Unilever brand being more liked/preferred compared to the competitor brand was seen. A larger scale test might show significance. Comments from participants indicated a preference of the taste of the Unilever brand over the competitor brand, and that degree of saltiness could be the reason not getting an even better result. Many participants found the Unilever Swedish brand being too salty, which is not surprising since Finland was one of the first countries in the world to attempt to reduce the sodium intake of its population (He and MacGregor, 2009), why they might be used to less salty food compared to Swedish people.

Tea 1 - Sweden
All three own brands were significantly more liked/preferred compared to one out of two competitor brands. The second competitor brand scored similar to the own brands. Since all the own brands showed good results in competition to competitors no further tests are needed.

Tea 2 - Sweden
Both of the two own brands were significantly more liked/preferred compared to two out of three competitor brands. The third competitor brand was scoring similar to the own brands. Since all the own brands showed good results in competition to competitors no further tests are needed.

Tea 3 - Sweden
Indications towards the own brand being more liked compared to the competitor brand was seen. A larger scale test might show if the own brand is to be significantly more liked/preferred over the competitor.

Tea 4 - Finland

Weak indications towards the own brand being slightly more preferred/liked over the competitor brand was seen. A larger scale ranking test might show if the own brand could be "Best in test".

Tea 5 - Finland

No indications regarding differences in preference/liking could be seen between the brands. The results were too similar to interpret what would be the result in a larger scale test. Comments from participants pointed towards the own brand having a stronger mint flavour compared to the competitors, a quality that some participants appreciated and some did not. No further tests are needed.

Tea 6 - Finland

Indications towards the own brand, Tea A, being more preferred over the competitor brand, and some indications towards the own brand, Tea B, being less preferred compared to the competitor brand was seen. The own brand, Tea B got a median value in "the disliking part" of the hedonic scale, why further evaluations are needed. A larger scale ranking test might show if the own brand, Tea A, is superior to the competitor brand, and if the competitor brand is superior to the own brand, Tea B.

Tea 7 - Finland

The Hedonic test gave similar grades in liking for both brands. The ranking test however showed indications towards the competitor being preferred over the own brand. A larger scale ranking test might show significance. Nevertheless, a larger scale hedonic test would probably not show much since the products are ranked very similar in liking.

8.4. Results from evaluation of bad performing product – Suggestions on improvements on Soup 1 A

When compared to the competitor products, Soup A got a median value for overall opinion of 6,5 = in between "dislike slightly" and "dislike moderately". However, when performing a single-product test the median value of Soup A was 3 = "like moderately". This indicates that the participants do not dislike the soup, but when compared to the other brands it is less preferred and gets a less liked grade. According to JAR-scales two attributes could be improved; the amount of herbs could be moderately increased and the sizes of meat pieces slightly increased. The hedonic scales with added attributes indicated that the smell/aroma of the soup was slightly disliked.

9. Conclusions

Performance of internal small scale consumer tests, to compare products can be a good way to identify possibly poor performing products and also outstanding good performing products. In this way expensive big scale consumer tests can be limited. However, performance of internal small scale consumer test requires adjustments and compromises in the sensory practices, why consideration regarding the reliability of the results is important.
In the evaluations of 23 Unilever products, seven products showed potential of being best in test, and six products got results indicating they were less preferred/liked compared to the competitor. The results showed that significant differences in liking and preference can be seen in test groups of only ten participants.

10. Acknowledgements

I have really enjoyed performing this project at Unilever. It has been a great experience and opportunity to learn a lot. Great thanks to my supervisor Lotta Beckman for all encouragement and support. You have been wonderful! Thanks to Marie Rydèn that gave me this opportunity, found this project for me and all the efforts you made to help me adapt it to suit both the Unilever and SLU requests. Thanks for letting me participate in interesting meetings and the opportunity to go to Finland and perform tastings. Thanks to Ulla-Maija Laitinen that made me feel so welcome at your office in Finland. I also want to thank my supervisor Cornelia Witthöfft at SLU for helping me with the last touch with the report and also Behnaz Mazogi for the patience and help with the statistics.

11. References

Dijksterhuis, G. B. (1997). Multivariate Data Analysis in Sensory and Consumer Science. *Food & Nutrition Press Inc*, Trumbull, CT, USA

Earle, M., Earle, R. and Anderson, A. (2001). Food product development, Woodhead Publishing Limited, *Woodhead Publishing in Food Science and Technology*, Cambridge

El Dine, A. N. and Olabi, A. (2009). Effect of reference foods in repeated acceptability tests: testing familiar and novel foods using 2 acceptability scales. *Journal of Food Science*, 74, S97-S106

He, FJ., MacGregor, GA. A. (2009). Comprehensive review on salt and health and current experience of worldwide salt reduction programmes. *Journal of Human Hypertension*. 2009;23(6):363–384.

Jaeger, S. R., Jørgensen, A. S., Aaslyng, M. D and Wender, L. P. Bredie, (2008). Best–worst scaling: An introduction and initial comparison with monadic rating for preference elicitation with food products. *Food Quality and Preference* 19, 579-588. Elsevier

Jones, L. V., Peryam, D. R. and Thurstone, L. L. (1995). Development of scale for measuring soldiers food preferences. *Food Research*, 20, 512-520

Lawless, H. T. and Claassen, M. R. (1993). The central dogma in sensory evaluation. *Food Technology*, 47(6), 139-146

Lawless, H. T. and Heymann, H. (2010). *Sensory Evaluation of Food – Principles and Practices.* Second Edition, Springer New York Dordrecht Heidelberg, London

López Osorino, M. M. and Hough, G. (2010). Comparing 3-point versus 9-point Just-About-Right-scales for determining the optimum concentration of sweetness in a baverage. *Journal of Science*, Volume 25,Issue Supplement s1, page 1-17

Ludovic Koehl, Xianyi Zenga, Bin Zhoua, Yongsheng Ding (2007). Intelligent sensory evaluation of industrial products for exploiting consumer's preference, *Mathematics and Computers in Simulation* 77, p 522–530

Lundgren, B. (2000). *Handbok i Sensorisk Analys*, SIK-Rapport Nr 470, Institutet för livsmedel och bioteknik, Kompendiet Lindome

Meilgaard, M. C., Civille, G. V. and Carr, B. T. (2007). *Sensory Evaluation Techniques*, Fourth Edition, CRC Press, Taylor and Francis Group, USA

Resureccion, A. V. (1998). Consumer Sensory Testing for Product development. Aspen, Gaithersburg, MD

Rothman, L. and Parker, M. J. (2009). *Just-About-Right Scales: Design, Usage, Benefits and Risks.* ASTM Manual MNL36, ASTM International, Conshohocken, PA

Schutz, H. G. (1965). A Food Action Rating Scale for Measuring Food Acceptance. *Journal of Food Science,* Volume 30, Issue 2, page 365-374

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

Sensorisk Studiegruppe (2006) *Sensorisk analyse – Bedømmelse av næringsmidler*, Gyldendal Undervisning, Universitetsforlaget, Oslo

Stone, H. and Sidel, J. L. (1993). Sensory Evaluation Practice, Academic Press Inc

Tuorila, H., Huotilainen, A., La¨hteenma¨ki, L., Ollila, S., Tuomi-Nurmi, S. and Urala, N. (2008). Comparison of affective rating scales and their relationship to variables reflecting food consumption. *Food Quality and Preference* 19, 51-61. Elsevier

Villanueva, N. D. M. and Da Silva, M. A. A. P. (2009). Performance of the nine-point hedonic, hybrid and self adjusting scales in the generation of internal preference maps. *Food Quality and Preference*, 20, 1-12

Appendix 1. Questionnaires

1.1. Example of questionnaire used at competitor tests

Competitor Tasting

Welcome and thank you for participating

You will get a number of samples containing different competitor products. The tasting consists of two parts/types of evaluation. Please read the instructions carefully and answer the questions. Keep in mind that you are asked to answer the questions as a representative of the consuming population; it is your personal opinion of liking and preference that is of interest. Since we want your personal opinion, please do not talk to the other participants during the test. Along with each question there is room for comments. Use this room to try to explain the reason to your choice as detailed as possible.

To help you reset your taste buds in between the samples, there are water and crisp bread available. Spiting after tasting is optional, but keep in mind that getting too full can change your perception of the products tasted closer to the end of the tasting session.

The product of today's tasting:

You will try **four brands of ……**

When performing consumer tastings it is preferable to let the consumer eat the product in the way that they normally do. In this case it's assumed that ……. often is eaten together with ……., why ……. is available as an optional accompaniment at the tasting. However, since ……. is strong smelling and tasting, make sure you try all the samples alone as they are before adding the ……..
.

The 9-point hedonic scale/The degree of liking scale

As a representative of the consuming population, quantify the degree of liking or disliking of the products one by one separately. Please try them from left to right, in the order presented. Put a cross in the box that best describes your overall opinion of the sample. Don't forget to rinse your mouth with water in between the samples.

Sample number __________

- ☐ Like extremely
- ☐ Like very much
- ☐ Like moderately
- ☐ Like slightly
- ☐ Neither like nor dislike
- ☐ Dislike slightly
- ☐ Dislike moderately
- ☐ Dislike very much
- ☐ Dislike extremely

Sample number __________

- ☐ Like extremely
- ☐ Like very much
- ☐ Like moderately
- ☐ Like slightly
- ☐ Neither like nor dislike
- ☐ Dislike slightly
- ☐ Dislike moderately
- ☐ Dislike very much
- ☐ Dislike extremely

Sample number __________

- ☐ Like extremely
- ☐ Like very much
- ☐ Like moderately
- ☐ Like slightly
- ☐ Neither like nor dislike
- ☐ Dislike slightly
- ☐ Dislike moderately
- ☐ Dislike very much
- ☐ Dislike extremely

Sample number __________

- ☐ Like extremely
- ☐ Like very much
- ☐ Like moderately
- ☐ Like slightly
- ☐ Neither like nor dislike
- ☐ Dislike slightly
- ☐ Dislike moderately
- ☐ Dislike very much
- ☐ Dislike extremely

Preference ranking test

As a representative of the consuming population, please taste the samples from left to right, in the order presented, and rank them from most preferred to least preferred (1= most preferred, 4= least preferred) You are allowed to re-taste the samples after trying them all. Remember to rinse your mouth with water in between the samples.

N.B. You will have to make a decision, ties are not allowed. (However if you find it hard to rank the samples please note it along with your comments.)

Ranking (1-4)	Sample	Comments
1	_____	_____
2	_____	_____
3	_____	_____
4	_____	_____

References and inspiration to the manuals

Harry T. Lawless & Hildegarde Heymann (2010). *Sensory Evaluation of Food*, 2nd ed. Springer-Verlag New York Inc, New York, NY

Lundgren Borgit (2000). *Handbok i Sensorisk Analys*, Svenska Livsmedelsinstitutet, SIK-Rapport Nr 470. Kompendiet-Lindome, Sverige

Personal data

	Male	Female
Gender:	☐	☐

Age group:	20-30	☐
	31-40	☐
	41-50	☐
	51-60	☐
	61-70	☐

	Yes	No
Do you in general like ……..?	☐	☐

In the last three month, about how often have you used the type of product of today's tasting?

☐	Not a single time
☐	Less than once a month
☐	More than once a month, but less than once a week
☐	More than once a week

Comments:

Please give your comments on the tasting session, the questionnaires etc. In that way improvements can be made.

1.2. Example of questionnaire used at further evaluation of poor performing product

Tasting

Welcome and thank you for participating

You will get one single sample/product. The tasting consists of two parts/types of evaluations. Please read the instructions carefully and answer the questions. Keep in mind that you are asked to answer the questions as a representative of the consuming population; it is your personal opinion of liking and preference that is of interest. Since we want your personal opinion, please do not talk to the other participants during the test. Along with each question there is room for comments. Use this room to try to explain the reason to your choice as detailed as possible.

The product of today's tasting:

You will try **one brand of ……..**

The 9-point hedonic scale/The degree of liking scale

As a representative of the consuming population, quantify the degree of liking or disliking of the product. Evaluate each given attribute one by one separately. Put a cross in the box that best describes your opinion of the product. Please try to give the reasons to your opinion under comments.

Appearance

□ Like extremely
□ Like very much
□ Like moderately
□ Like slightly
□ Neither like nor dislike
□ Dislike slightly
□ Dislike moderately
□ Dislike very much
□ Dislike extremely

Comment your choice:

_____

_____

Smell/aroma

□ Like extremely
□ Like very much
□ Like moderately
□ Like slightly
□ Neither like nor dislike
□ Dislike slightly
□ Dislike moderately
□ Dislike very much
□ Dislike extremely

Comment your choice:

_____

_____

Taste

□ Like extremely
□ Like very much
□ Like moderately
□ Like slightly
□ Neither like nor dislike
□ Dislike slightly
□ Dislike moderately
□ Dislike very much
□ Dislike extremely

_____

_____

Texture

□ Like extremely
□ Like very much
□ Like moderately
□ Like slightly
□ Neither like nor dislike
□ Dislike slightly
□ Dislike moderately
□ Dislike very much
□ Dislike extremely

_____

_____

Over all opinion

□ Like extremely
□ Like very much
□ Like moderately
□ Like slightly
□ Neither like nor dislike
□ Dislike slightly
□ Dislike moderately
□ Dislike very much
□ Dislike extremely

_____

_____

Just About Right Scales

As a representative of the consuming population, please give your opinion on the specific attributes mentioned. Evaluate each given attribute one by one separately. Put a cross in the box that best describes your opinion of the product. Please try to give the reasons to your opinion under comments.

Saltiness

□ Very much too salt
□ Too salt
□ Slightly too salt
□ Just about right
□ Slightly not salt enough
□ Not salt enough
□ Very much not salt enough

Comment your choice:

_____

_____

Sweetness

□ Very much too sweet
□ Too sweet
□ Slightly too sweet
□ Just about right
□ Slightly not sweet enough
□ Not sweet enough
□ Very much not sweet enough

Comment your choice:

_____

_____

Thickness/thinness

□ Very much too thick
□ Too thick
□ Slightly too thick
□ Just about right
□ Slightly too thin
□ Too thin
□ Very much too thin

Comment your choice:

_____

_____

Herbs

□ Very much too much herbs
□ Too much herbs
□ Slightly too much herbs
□ Just about right amount of herbs
□ Slightly not enough herbs
□ Not enough herbs
□ Very much not enough herbs

Comment your choice:

_____

_____

Meat pieces (amount)

□ Very much too many
□ Too many
□ Slightly too many
□ Just about right
□ Slightly too few
□ Too few
□ Very much too few

Comment your choice:

_____

_____

Meat pieces (size)

□ Very much too big
□ Too big
□ Slightly too big
□ Just about right
□ Slightly too small
□ Too small
□ Very much too small

Comment your choice:

_____

_____

Personal data

	Male	Female
Gender:	☐	☐

Age:	20-30	☐
	31-40	☐
	41-50	☐
	51-60	☐
	61-70	☐

	Yes	No
Do you in general like ……..?	☐	☐

Comments:

Please give your comments on the tasting session, the questionnaires etc. In that way improvements can be made.

Appendix 2 – Example of detailed results from competitor tests

RESULTS – Soup 1, Sweden – 27/2-2012

Evaluation of: Soup A (Unilever), Soup B, Soup C and Soup D

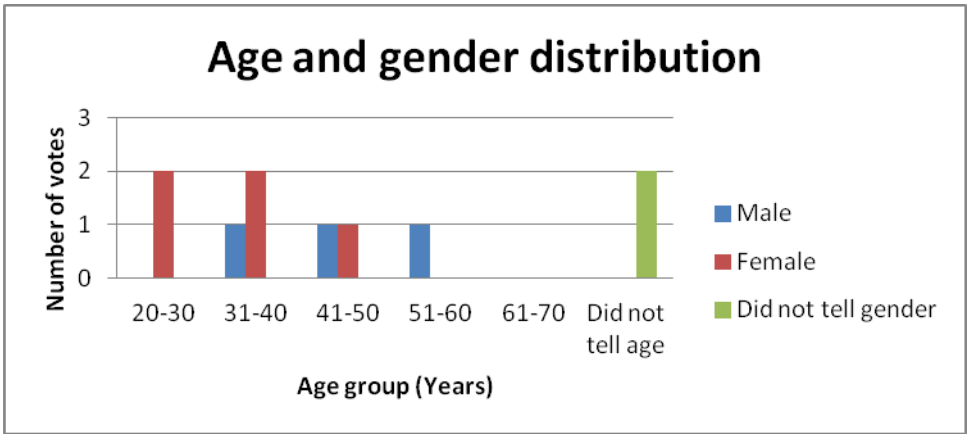

Figure 1. 10 individuals participated in the hedonic test and the ranking test; five females, three males and two that did not tell gender.

Hedonic 9-piont scale test

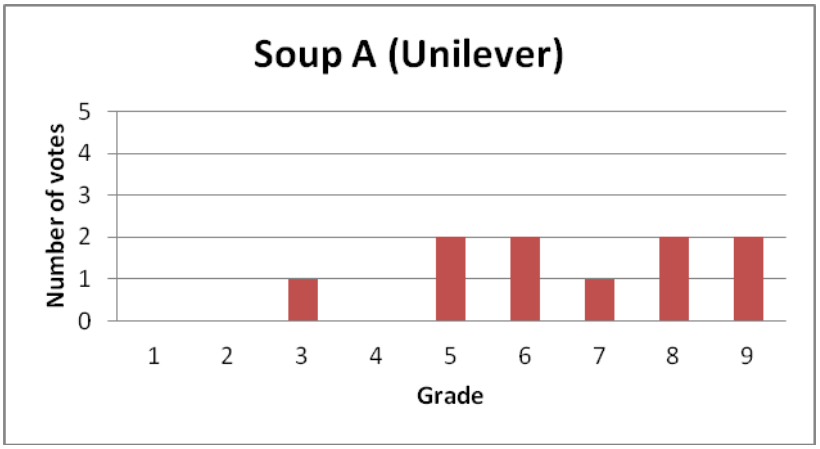

Figure 2. Hedonic 9-point scale grading of Soup A (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in the right part of the diagram, i.e. “the disliking part”.

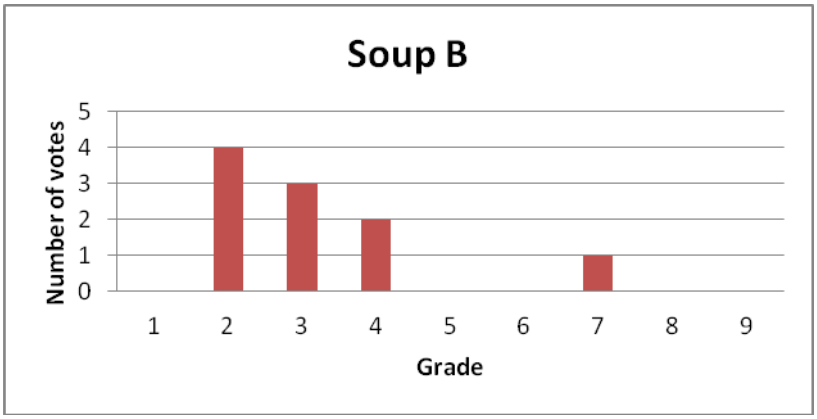

Figure 3. Hedonic 9-point scale grading of Soup B (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in the left part of the diagram, i.e. “the liking part”.

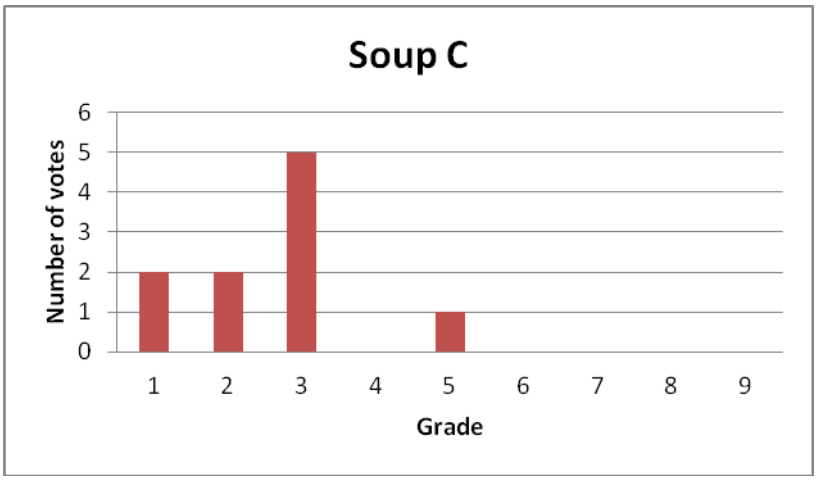

Figure 4. Hedonic 9-point scale grading of Soup C (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in the left part of the diagram, i.e. “the liking part”.

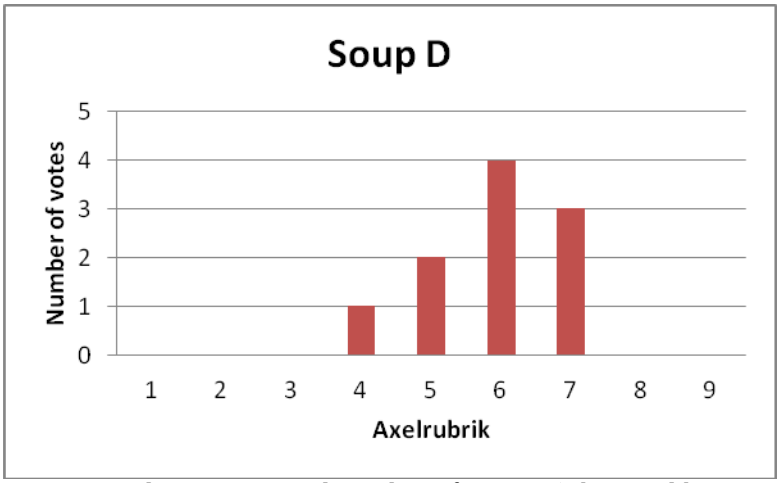

Figure 5. Hedonic 9-point scale grading of Soup D (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in the right part of the diagram, i.e. “the disliking part”.

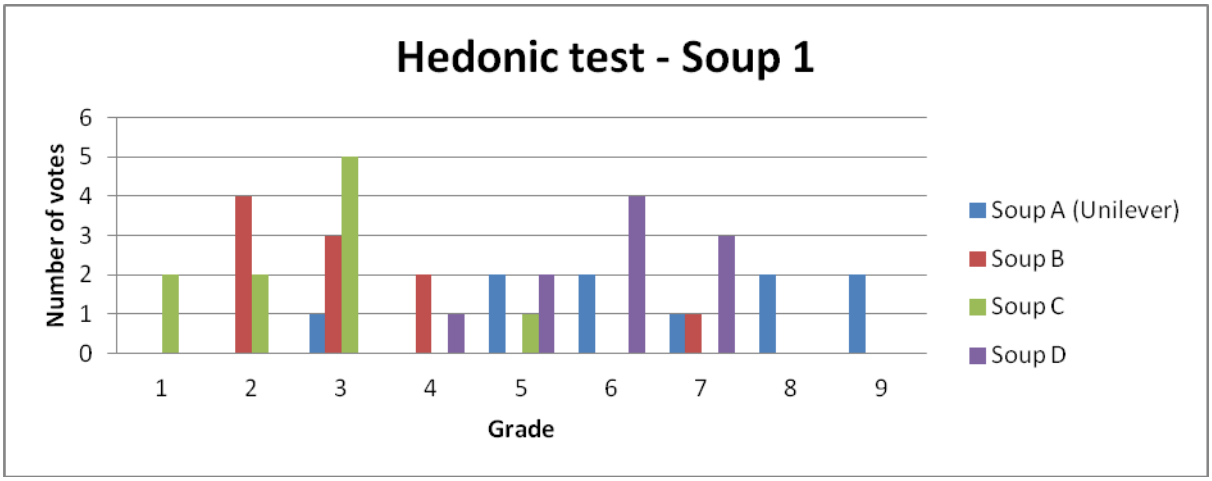

Figure 6. Hedonic 9-point scale grading of Soup A, B, C and D (where 1= like extremely and 9 = dislike extremely). The figure shows that the majority of votes for Soup B and C are located in the left part of the diagram, i.e. "the liking part", while the majority of votes for Soup A and D are located in the right part of the diagram, i.e. "the disliking part".

Table 1. The table shows that Soup B and C got the same median value (3= "like moderately"), while Soup A and D got almost the same median values (6 and 6,5 ="dislike slightly" and in between "dislike slightly" and "dislike moderately"). The T-test shows that Soup A is significant less liked compared to Soup B (p= 0,002), and Soup C (p=0,001). No significantly difference in liking between Soup A and D can be seen.

Brand	Soup A (Unilever)	Soup B	Soup C	Soup D
Participants	10	10	10	10
Sum	66	32	26	59
Mean (average)	6,6	3,2	2,6	5,9
Standard deviation	1,955	1,549	1,174	0,994
Median value	6,5	3	3	6

Significance – The Paired T-test

Table 2. Critical t-values at 9 degree of freedom

Level of significance	0,2	0,1	0,05	0,02	0,01	0,002	0,001
P	1,383	1,833	2,262	2,821	3,250	4,297	4,781

Soup A – Soup B = 4,543
Significance p = 0,002 (Soup A is less liked compared to Soup B)

Soup A – Soup C = 5,164
Significance p=0,001 (Soup A is less liked compared to Soup C)

Soup A – Soup D = 1,210
No significant difference in liking

Ranking test

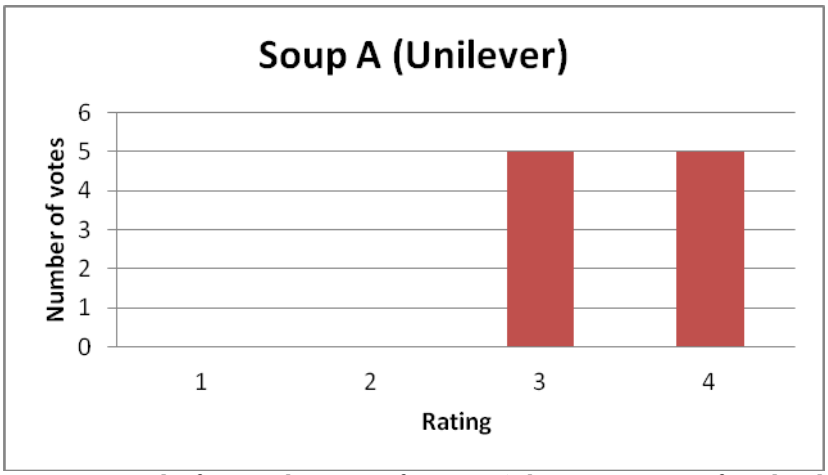

Figure 1. Results from ranking test of Soup A (where 1 = most preferred and 4 = least preferred). All votes are located in the right part of the diagram, i.e. “the less preferred part”.

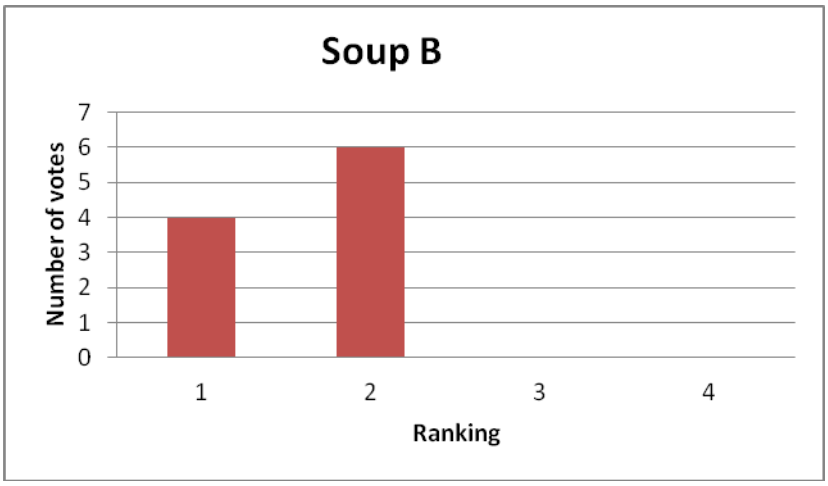

Figure 2. Results from ranking test of Soup B (where 1= most preferred and 4 = least preferred). All votes are located in the left part of the diagram, i.e. “the more preferred part”.

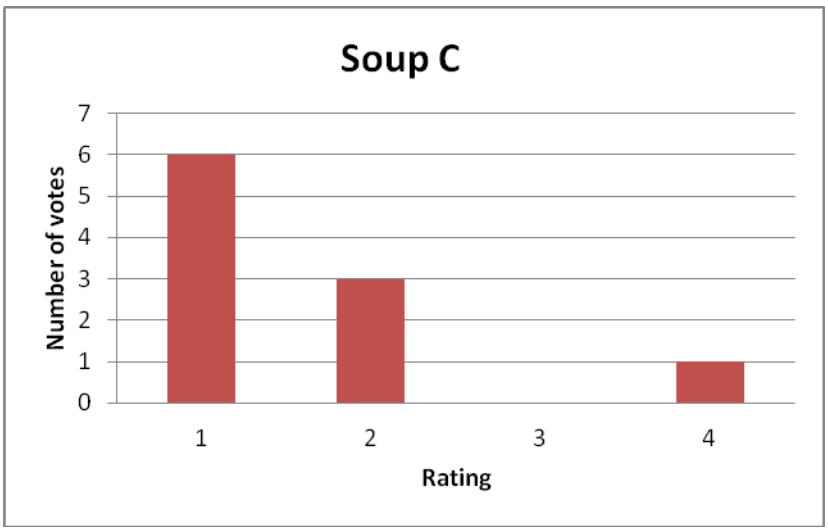

Figure 3. Results from ranking test of Soup C (where 1= most preferred and 4 = least preferred). The majority of the votes are located in the left part of the diagram, i.e. “the more preferred part”.

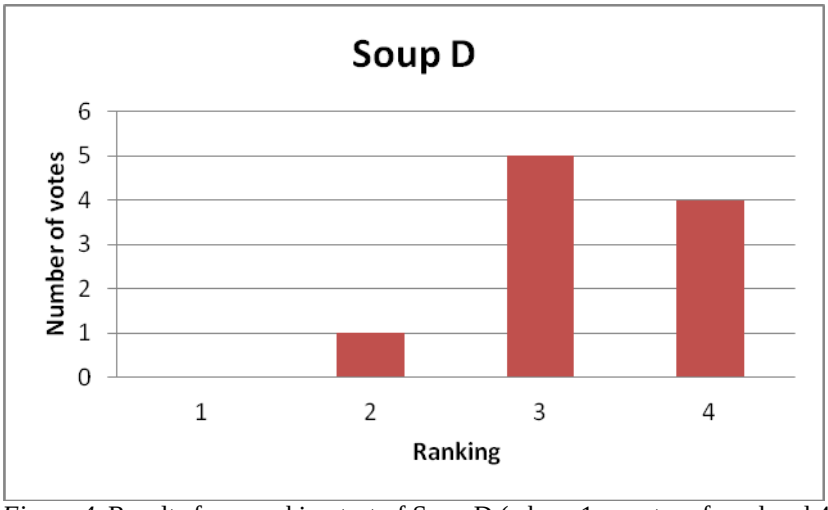

Figure 4. Results from ranking test of Soup D (where 1=most preferred and 4 = least preferred). The majority of the votes are located in the right part of the diagram, i.e. "the less preferred part".

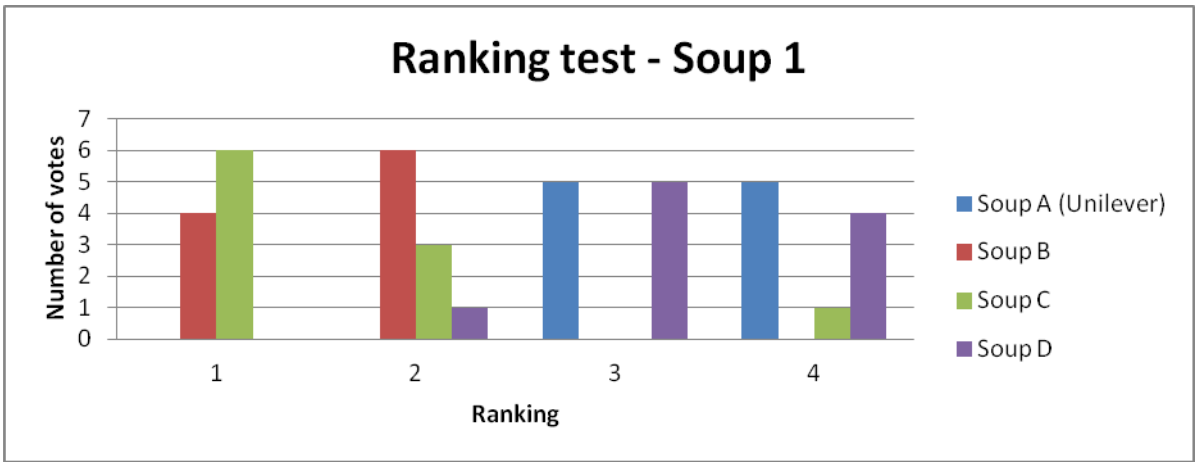

Figure 5. Results from ranking test of Soup A, B, C and D (where 1= most preferred and 4 = least preferred). The figure shows that almost all votes for Soup B and C are placed in the left part of the diagram ("the more preferred part"), and almost all votes for Soup A and D are placed in the right part of the diagram ("the less preferred part").

Table 1. The table shows that Soup C got the best median value, followed by Soup B. Soup A got the highest median value followed by Soup D. The Friedman test shows that there is a significant difference in preference between the brands (p=0,05), and LSD shows that Soup B and C are significantly more preferred over Soup A and D. However, no significant difference in preference between Soup A and D or between Soup B and C is seen.

Brand	Soup A (Unilever)	Soup B	Soup C	Soup D
Participants	10	10	10	10
Sum	35	16	16	33
Mean (average)	3,5	1,6	1,6	3,3
Standard deviation	0,527	0,516	0,966	0,675
Median value	3,5	2	1	3

Comments from participants:

Soup A (Unilever) – Nine out of ten participants had complains regarding the taste: does not taste fresh, tastes like old meat, poor quality meat, has a sub taste, bad taste, or it does not taste like pea soup. Two people disliked the consistency and found it to be jelly or more like a pure.
Soup B - Almost all the participants mentioned the taste as good. There were different opinions regarding the colour; some found it too yellow and some liked it. The majority liked the texture and mouth feel.

Soup C – The majority of participants liked the taste and the texture

Soup D – Eight participants mentioned negative aspects regarding the taste; strange taste, don't taste like …. soup, after taste or too little taste. Six people thought it was too thin, and three people did not like the appearance of the soup.

Appendix 3 – Results from further evaluation of soup 1A

Follow up – Soup 1 26/4 – 2012

Soup A (Unilever)

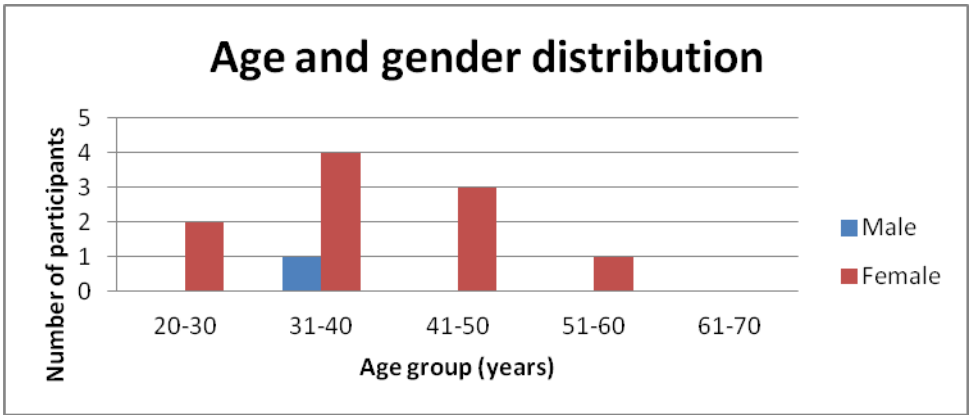

Figure 1. Eleven people participated in the test; ten females and one male. (However, in the JAR-scale question regarding size of meat pieces one female age 31-40 did not answer).

Hedonic 9-piont scale test – with added attributes

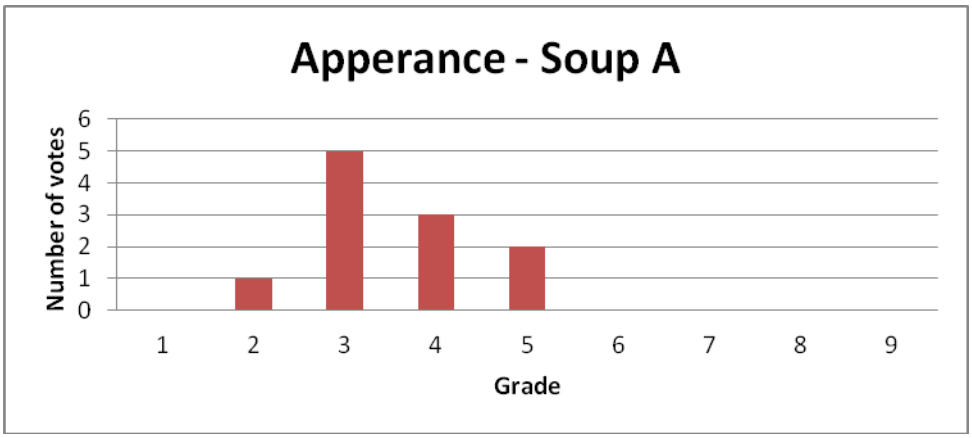

Figure 2. Hedonic 9-point scale grading of appearance of the soup (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in "the liking part" of the diagram.

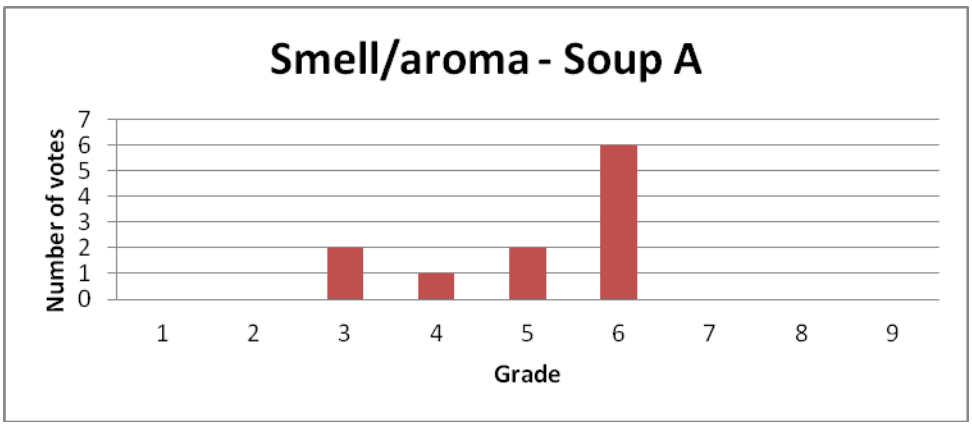

Figure 3. Hedonic 9-point scale grading of smell/aroma of the soup (where 1= like extremely and 9 = dislike extremely). The opinions regarding the smell/aroma differ among the participants. However, the majority of votes are located in "the disliking part" of the diagram.

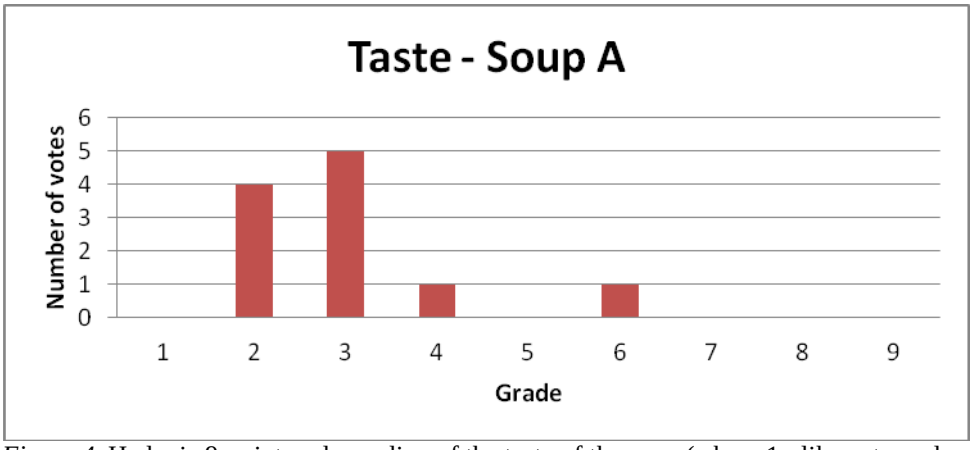

Figure 4. Hedonic 9-point scale grading of the taste of the soup (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in "the liking part" of the diagram.

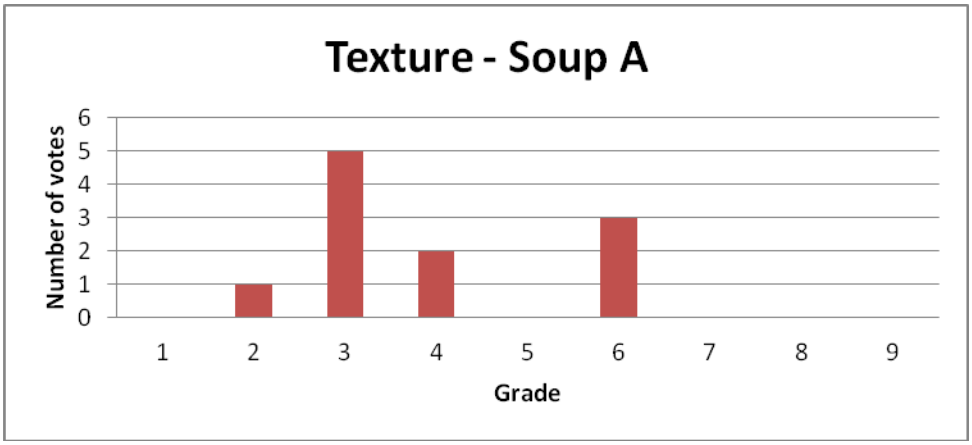

Figure 5. Hedonic 9-point scale grading of the texture of the soup (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in "the liking part" of the diagram.

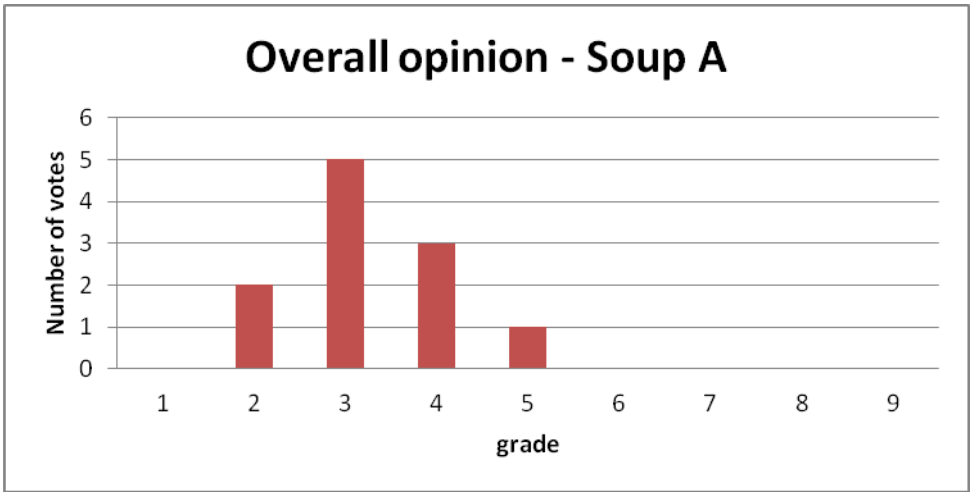

Figure 6. Hedonic 9-point scale grading of the overall opinion of the soup (where 1= like extremely and 9 = dislike extremely). The majority of votes are located in "the liking part" of the diagram.

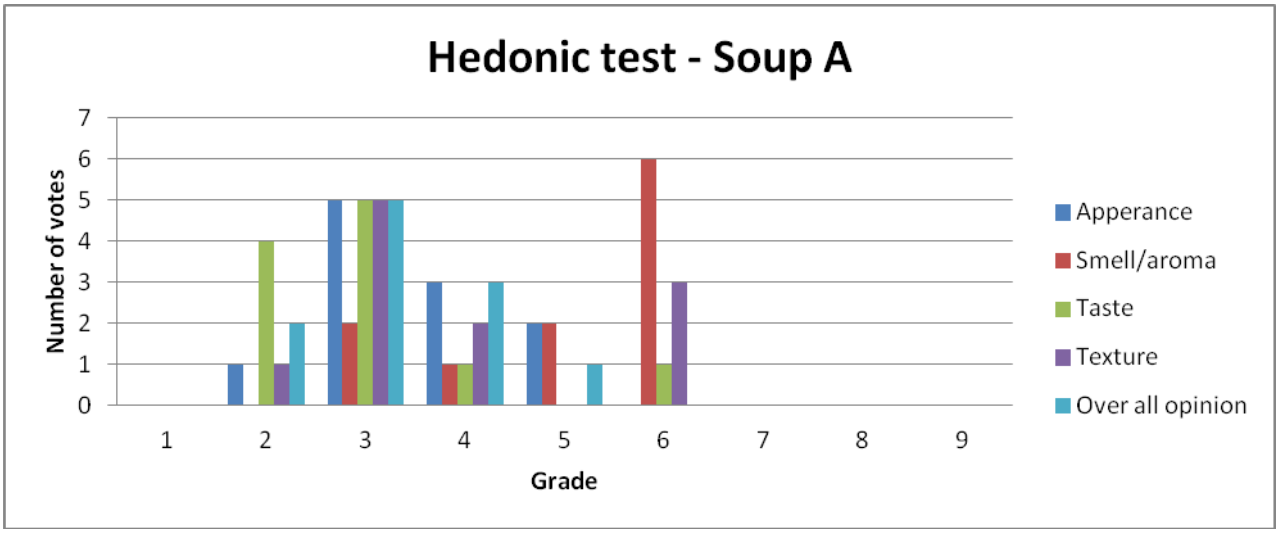

Figure 7. Hedonic 9-point scale grading of the appearance, smell/aroma, taste, texture and overall opinion of the soup (where 1= like extremely and 9 = dislike extremely). The majority of votes for all attributes except for smell/aroma are located in "the liking part" of the diagram. Smell/aroma however has the majority of votes located in "the disliking part" of the diagram.

Table 1. The taste got the best mean value, followed by overall opinion, appearance, texture and last smell/aroma. All attributes, except for "smell/aroma" got a median value of 3 (="like moderately"). "Smell/aroma" however got a median value of 6 (="dislike slightly").

	Appearance	Smell/aroma	Taste	Texture	Overall opinion
Participants	11	11	11	11	11
Sum	39	56	33	43	36
Mean	3,545	5,091	3	3,909	3,273
Standard deviation	0,934	1,221	1,183	1,446	0,905
Median value	3	6	3	3	3

Just-About-Right- Scales

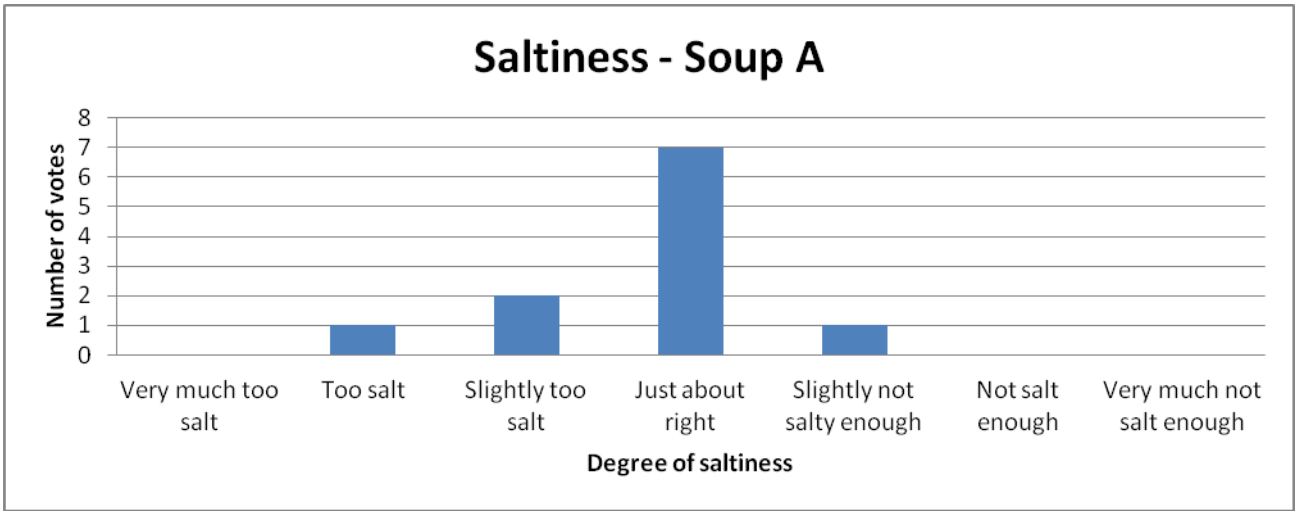

Figure 1. JAR-scale of saltiness of soup A. The majority of votes are located at "just about right".

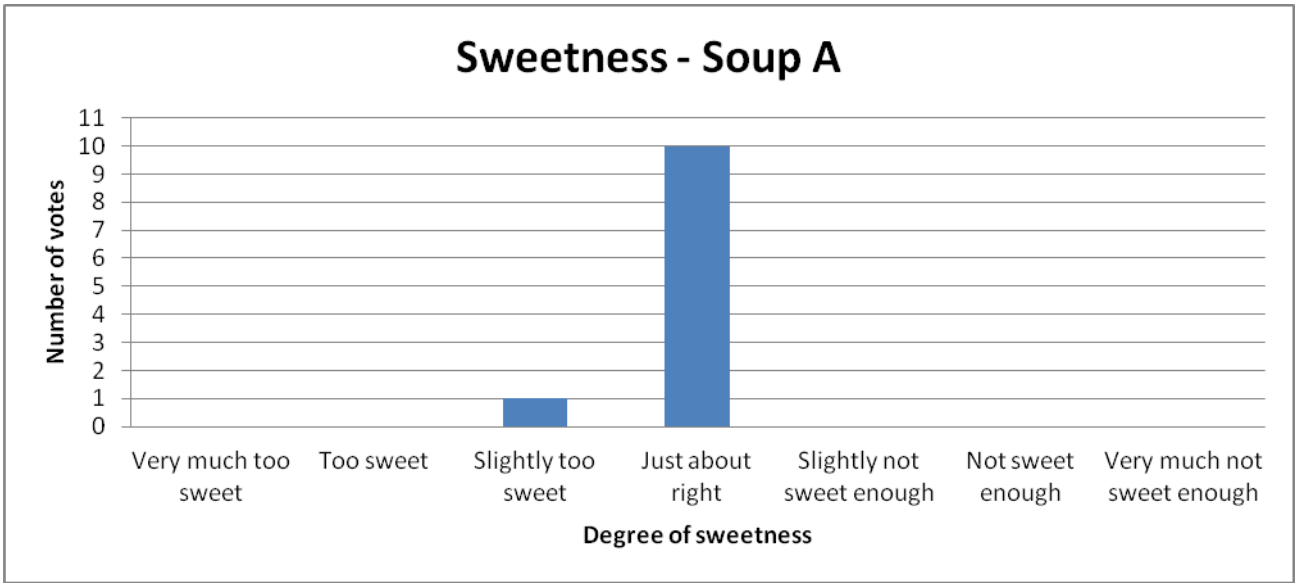

Figure 2. JAR-scale of sweetness of soup A. The majority of votes are located at "just about right".

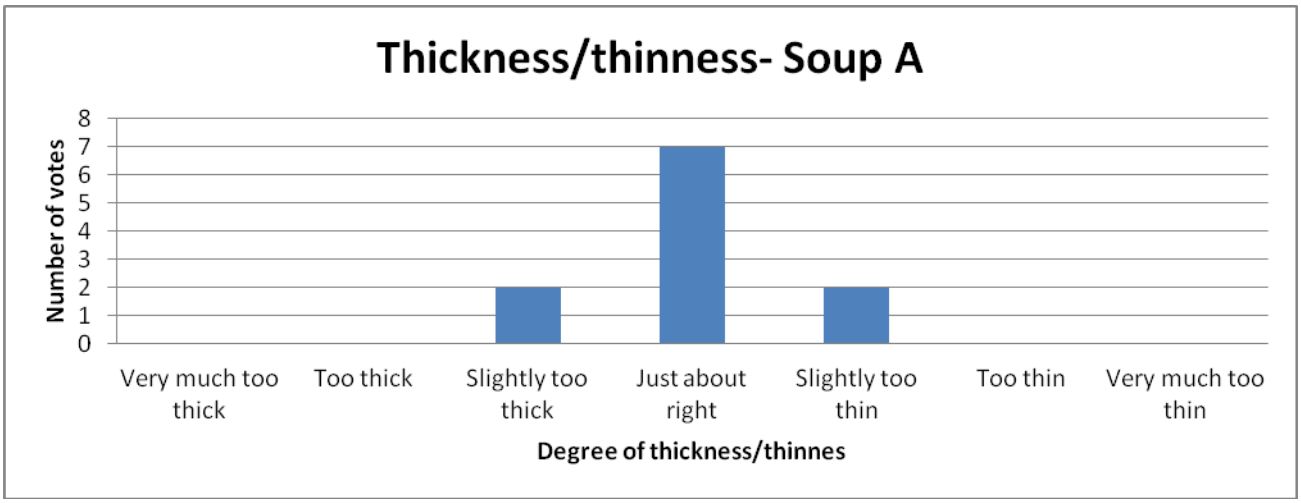

Figure 3. JAR-scale of thickness/thinness of soup A. The majority of votes are located at "just about right".

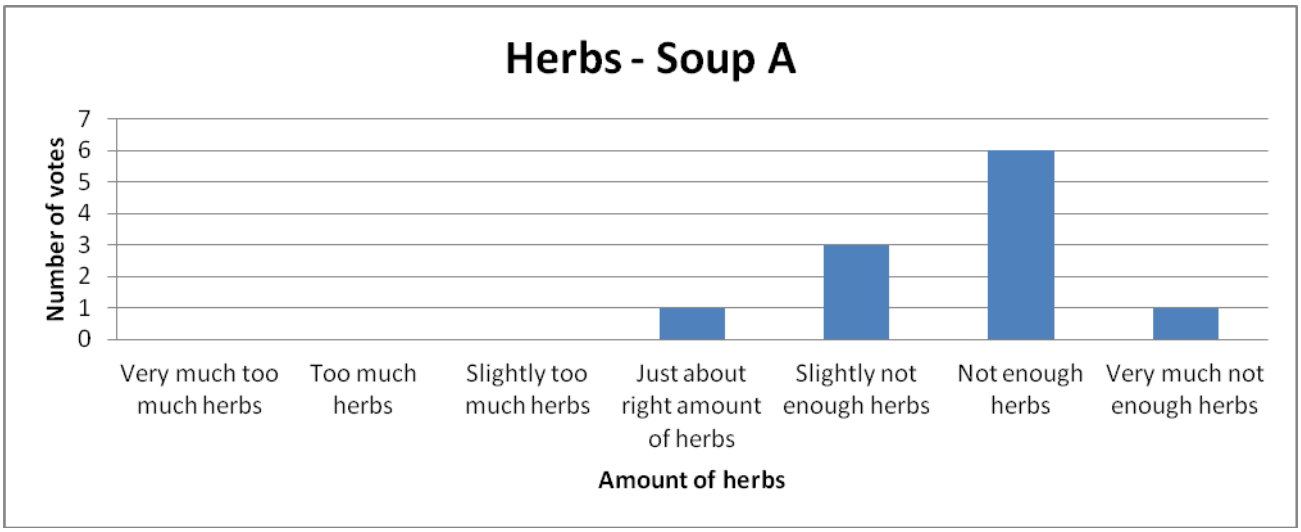

Figure 4. JAR-scale of amount of herbs in soup A. The majority of votes are located in the right part of the diagram, i.e. in "the not enough herbs part".

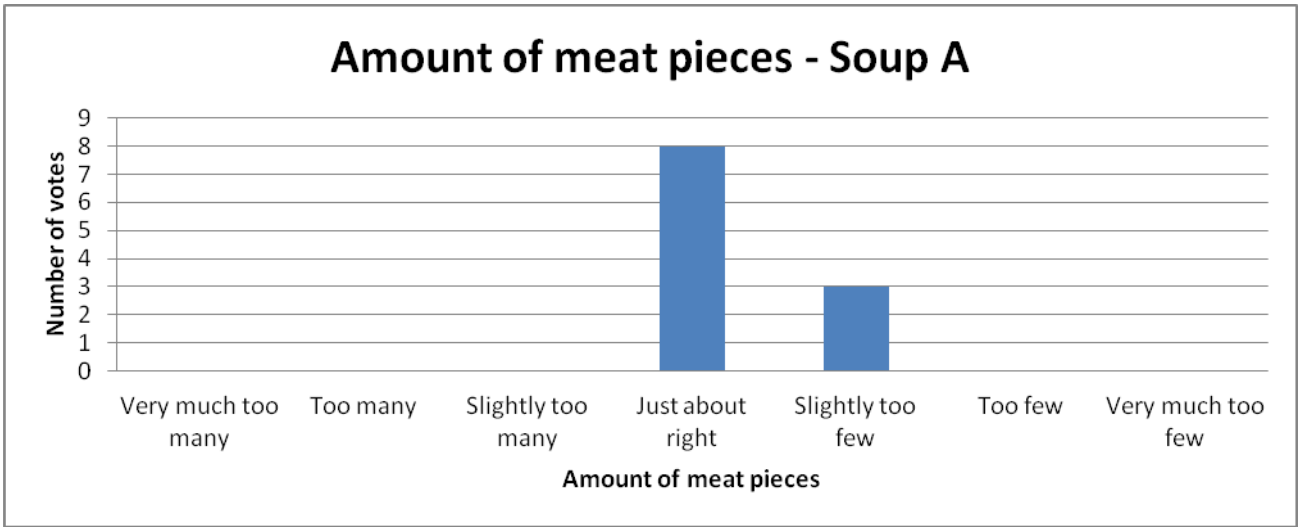

Figure 5. JAR-scale of amount of meat pieces in soup A. The majority of votes are located at "just about right".

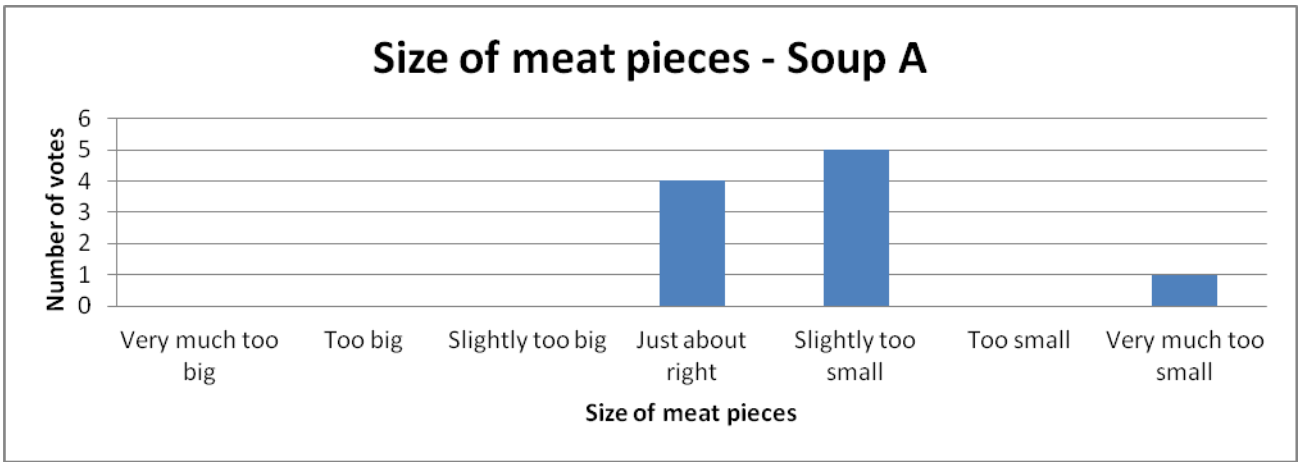

Figure 6. JAR-scale of size of meat pieces in soup A. The majority of votes are located in the right part of the diagram, i.e. in "the too small part". However, there are quite many votes on "just about right" too.

Table 1. Thickness got the best mean value (closest to 4="just about right"), followed by Sweetness, Meat pieces (amount), saltiness, meat pieces (size) and last herbs..

	Saltiness	Sweetness	Thickness	Herbs	Meat pieces (amount)	Meat pieces (size)
Participants	11	11	11	11	11	10
Sum	41	43	44	62	47	48
Mean value	3,727	3,909	4	5,636	4,273	4,8
Standard deviation	0,786	0,302	0,632	0,809	0,467	0,919
Median value	4	4	4	6	4	5

Comments from participants

Appearance:
Pieces of pork fat visible - not pleasant
Nice to see pieces of meat
Boring color, could be little bit less brownish, the color can be more bright, looks grayish
I like the color, nice and yellow color
Too smooth – the product is mashed

Smell/aroma
This kind of soup does in general not smell good
It doesn't smell that much at all
Difficult to tell just by the smell that this is that kind of soup
Bland smell
Where is the thyme
No herbs or other spices and no meat smell
Smells greasy

Taste
I like that there isn't much herbal taste
I like the flavor from the meat
Ok taste, good taste, nice taste
Miss herbs
Not enough salt
too salt x2

Texture
Too crushed x5
Missing large pieces of meat
A bit too fluid

Over all opinion
Ok, no really distinct taste
Over all good, good soup
Good distribution of ingredients
More herbs please!

Appendix 4 – List off relevant attributes for the different food categories for further evaluation of poor performing products

4.1. Added attributes for the hedonic liking scale – All product types

- Appearance
- Smell/odor
- Texture
- Taste
- Over all opinion

4.2. Attributes for JAR-scales - Tea

- Sweetness
- Bitterness
- Harshness
- Strong/weak flavor
- Fruitiness
- Perfume taste
- Colour – too dark/too light
- sour/acidity

4.3. Attributes for JAR-scales - Soup

- Sweetness
- Saltiness
- Strong/weak flavor
- Amount of herbs, spices
- Sour/acidity
- Thickness/thinness
- Amount of meat pieces or vegetable pieces
- Size of meat pieces or vegetable pieces
- Softness/hardness on vegetables

4.4. Attributes for JAR-scales - Margarine/butter

- Sweetness
- Saltiness
- Hardness/softness

Appendix 5 – Popular scientific summary of the report

Consumer tests can provide a company with important and useful information regarding both sensory characteristics (i.e. taste, smell, appearance, consistency/texture etc.) of their products and information regarding the consumer liking and preferences (Lawless and Heymann, 2010). This information is crucial in determining and maintaining the quality of a product, in the work towards new product development, in the forecasting of market behavior and when exploiting new markets (Koehl et al., 2007). However, performance of large scale consumer tests are often very expensive, why alternative approaches and new cost-effective options are constantly developed (Meilgaard et al., 2007). This report is part of a new sensory project at Unilever. It presents a method for performance of small scale internal consumer tests that allows sensory comparison of the company's own products with the corresponding competitor products. The method gives the company a cheap way to get valuable information regarding their products advantages and flaws, and allows them to identify possibly poor performing products and also outstanding good performing products. In this way expensive big scale consumer tests can be limited into only including products that likely needs improvements or that in a larger scale tests could be proven "best in test". Out of many potential tests for affective evaluation of foods two tests were chosen; the 9-point hedonic test and the ranking test. The 9-point hedonic test evaluates the liking of a product, while the ranking tests gives answers to which product that is most/least preferred. Questionnaires and preparation procedures were constructed, where after 16 evaluations including 23 Unilever products, performed on different food categories, were conducted. The results showed that indications and also significant differences in liking and preference could be seen in test groups of only ten participants. Among the evaluations performed seven Unilever products showed potential of being best in test, and six Unilever products got results indicating they were less preferred/liked compared to the competitor, why further evaluations are needed.

As an example of a possible way to proceed with the identified poor performing products, to find out what attributes that may have caused the bad outcome, a second method was presented and practiced on one product. The results showed that a product can be poor performing and graded as disliked among participants when compared to competitor products, but when tested on its own regarded as acceptable or even good performing. Further evaluations will have to be performed before determination if this is a successful method of identifying which attributes that causes the product flaws.

This report has shown that performance of internal small scale consumer tests, to compare products, can be a good way to at low costs identify possibly poor performing products and also outstanding good performing products. However, performance of internal small scale consumer tests requires adjustments and compromises in the sensory practices, why consideration regarding the reliability of the results is important. Nevertheless, since the purpose of the results of the small scale internal tests are not to be used as a proof, but only as a tool to help identify products that for some reason might be in need for further evaluations, the need for precise evaluation conditions and practices are not as strictly essential. It will be up to the company to decide to what extent they believe that the results are reliable.