DANIELLE VAN HOUT

Measuring Meaningful Differences

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

Measuring Meaningful Differences

Measuring Meaningful Differences

Sensory testing based decision making in an industrial context; applications of signal detection theory and Thurstonian modelling

Belissingen maken op basis van sensorisch onderzoek in een industriele context; toepassingen van signaal detectie theorie en Thustoniaanse modellen

THESIS

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.

The public Defence shall be held on
Thursday 23rd of January 2014

By

Danielle Helena Alice van Hout
born in Kerkrade

ERASMUS UNIVERSITEIT ROTTERDAM

Doctoral Committee:

Promotors: Prof.dr. P.J.F. Groenen
Prof.dr. G.B. Dijksterhuis

Other members: Prof.dr. A.R. Thurik
Prof.dr. R.A. Zwaan
Prof.dr.ir. C. de Graaf

Erasmus Research Institute of Management – ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM PhD Series in Research in Management, 304
ERIM reference number: EPS-2014- 978-90-5892-350-9
ISBN 978-90-5892-350-9

Cover illustration: Down the Rabbit Hole (1969) from Alice in Wonderland Suite, by Salvador Dali
Design: B&T Ontwerp en advies www.b-en-t.nl

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

ACKNOWLEDGEMENTS

Looking back over the last couple of years since I began "working on my PhD", I slowly started to realize what it all involved. Since then it has been quite a journey for me to reach this point. There have been many obstacles and challenges, but also new opportunities. Now that I am writing these thank-you words for all these great and special people around me, I seem to get more used to the idea that I made it to the finish! There have been many people on my path that have been instrumental to this achievement and I would like to spend some words of thanks for their support.

In the first place I want to thank my promoters Patrick Groenen and Garmt Dijksterhuis for their coaching and support, their contributions, and their patience to let me develop the necessary knowledge and skills for bringing this to a successful end. I especially appreciated your willingness to dive deep into this science field which resulted in the many good discussions that we had. This all made "working on my PhD" a very valuable experience.

This research has all been part of my work in Unilever and I would like to thank my company for the great opportunity that I could do this. In particular, the support I got in terms of freedom to work on this topic, and our mutual view that this topic is important. There are a few colleagues that I would like to mention specifically. First of all I would like to thank my good colleague and friend Anna Thomas. Anna was the one who encouraged, and convinced me to take this step as it would be a great personal development for me. I can say for sure that without Anna I never would have started this journey. A special thank you as well for Liesbeth Zandstra who regularly took on the role of coach in terms of time management, reflecting on the importance of keeping this research a priority amongst the various priorities in my daily work, as well as for her valuable advice on the content of this work. Next to this, I wish to thank my colleagues in "the Vlaardingen lab", and my colleagues in the different countries with whom I have been working together over the years. A final word of thanks goes to my old SPB Perception team for their great support, and for remaining interested about the research and including it into their work with great enthusiasm.

Then, there are a few very special people in the academic world that have played major roles in my enthusiasm for this specific topic, and in my curiosity to keep on studying it in more and more detail. First of all there is Dr. Michael O'Mahony, Professor of sensory science at the University of California, Davis. He has been the one who introduced the concepts of signal detection theory and Thurstonian modelling, and the importance of this for sensory and consumer science, to the broader audience of sensory practitioners. Mike O'Mahony is one of the most original thinkers in the field, and to my opinion he is also the best speaker in the world. Since I attended a short course from Mike in 1995, this topic never left my mind again and for most of the research I was involved in from this point onwards, I continued in one way or the other to build further on it. The second person I would like to thank is Dr. Michael Hautus, Associate Professor of Psychophysics at the Auckland University, New Zealand. Michael and I have been working together in research projects for many years and I learned a lot about signal detection and psychophysics from him. I experienced how the psychological terminologies are different from how things are called in sensory science, and his knowledge on the different sensory modalities provided many new ideas to research further. Finally, last but definitely not least, I would like to say a special word of thanks to my dear friend Dr. Hye-Seong Lee, Associate professor of Sensory Science and Ergonomics at EWHA University, Seoul, South Korea for her enthusiasm, huge energy and support during the more than 10 years that we have been doing research together. Hye-Seong, I think it's fantastic to know you and that I had this chance of developing ideas together with you, learning about the differences between industry and academic research, and also experiencing the cultural differences – and how surprisingly similar some things turn out to be. Together, we identified many issues in sensory and consumer test methods, discussed solutions, and designed experiments that often have led to valuable insights. I hope that we will keep continuing our research together for many, many years and look very much forward to seeing your excellent students develop into successful scientists with a sunny future. They are very lucky with a professor like you.

For the next people I would like to thank, I can say for sure that without them, we literally wouldn't have been here now. Therefore very special thanks to my parents, not only for welcoming me to this world and for guiding me with their parental coaching while I grew up, but in particular for the way in which they did this. I enormously valued that they always listen to understand my point of view, and give me the freedom to do things in my own way. I think that this has been essential for me to reach this point. Further thanks to my family and friends for their support during these years. Thank you, Reijer, for our discussions on the topic and for believing that I could (easily...) do this, and for making me believe this too (Well, it was actually not so easy, but I made it!) Thanks as well to my parents and Annie for proof reading the manuscript, and special thanks to my sister and friend Pauline, for helping me to get the subject clear into my mind and helping me learn to like academic writing.

Finally, I want to thank my partner Wilfred for his huge support through these years and for standing with me. I had to spend a lot of "our time" on this PhD project, time that we could potentially do fun things together. Instead of keeping distant, you started to learn more and more about my research and we could have discussions together, so this became shared fun as well. Our discussions helped me a lot to make things clearer in my head as well as with communicating about it. In our day to day lives, it was also excellent that you could take so many things out of my hands, so that I could focus on writing my PhD.

Many thanks to all of you! I am a very lucky person to know you

Contents

1. INTRODUCTION

1.1 Challenges for sensory and consumer testing in a FMCG environment

In the ‘fast moving consumer goods’ (FMCG) industry, results from sensory product tests form the basis for many important business decisions; decisions on whether or not to launch new products, change existing products, or whether to continue with specific novel technological developments. It is therefore important that sensory tests are accurate and deliver robust results.

In most corporate functions where decision making takes place, ‘action standards’ are defined, based on direct measures of effect size. For example, in marketing whenever a new innovation is launched, post-launch marketplace performance measures are available, such as sales and complaints data, and in supply chain functions there is accurate information on logistics and pricing of raw materials, which makes it possible to make informed decisions.

For sensory product testing, direct measures of effect size are usually not possible, as these tests typically take place during the very early stages of development, long before the products enter the market. This means that we need to rely on indirect measures using human subjects, to predict potential in-market performance. The main problem with tests involving with human subjects is the noise in the measurement, caused by the various internal and external influences, both physiological and cognitive in nature, affecting test performance.

The conventional statistical approaches for sensory data analysis, sometimes called deterministic models, or guessing models, are on their own not very well suited to modelling human data. The main reason is that statistical approaches do not take into account the physiological and cognitive factors in tests that can influence test performance. Consequently, sensory results such as the presence or absence of a difference between products depend on the method that is used. Due to this dependency, it is not possible to directly compare results from several sensory studies and gain a deeper understanding on effect sizes. Such understanding can be very valuable for a company as it will increase the predictability of the results, for example by calibrating differences ‘in-lab’ to ‘meaningful differences’ that consumers would notice when they use the product. This will allow for better decision making and will also make further testing more effective in time and costs.

1.2 Using Signal Detection Theory and Thurstonian modelling

The problem of method-specificity of sensory results can be solved by using an alternative type of data analysis and modelling in addition to the traditional statistical approach. The framework of Signal Detection Theory (Green and Swets, 1966) and Thurstonian modelling (Thurstone, 1927) makes it possible to investigate the internal and external factors in tests and study how these factors influence subjects' test performance. In particular, this framework can be used to study subjects' psychological and physiological processes that influence test performance; the *perceptual process* that integrates the information from the senses and the *decision process* that uses the integrated information and the task instructions to make a decision. Signal Detection Theory provides models of the relationships between different test methods, and enables the development of tools for studying and optimizing test performance. It has led to a standardised measure of sensory difference, called d-prime (*d´*), available for most used test types. The *d´* allows to select the best test method for specific purposes and to compare results between different test types, e.g. comparing in-lab test results to findings of consumer in-home tests, so that optimal Action Standards can be defined.

While the science has been around for a while, it has taken time until an understanding was built in psychophysics that allowed investigating the perceptual processes of the different senses. Signal Detection Theory approaches were further explored in academic psychological laboratories (O'Mahony, 1972a, b, c; Tedja, Nonaka, Ennis and O'Mahony, 1994; Frijters, 1979a, b; Irwin, Stillman, Hautus and Huddleston, 1993) using mainly simple stimuli. Sensory testing of real products is more complex (Hautus and Irwin, 1995; Rousseau, Rogeaux and O'Mahony, 1998). When Signal Detection Theory was introduced in sensory science (O'Mahony, 1972, 1979), first the focus was on exploring the factors influencing the *perceptual processes*, leading to more effective test designs (O'Mahony 1995b; O'Mahony and Godman, 1974). After this, the focus shifted towards understanding and optimizing the *decision processes* in tests (Hautus, O'Mahony and Lee, 2008; Hautus, Shephert and Peng 2011a, b), leading to the development of more effective tests that are more predictive of consumers' reality (Boutrolle, Delarue, Koster, Aranz and Danzart, 2009; Chae, Lee and Lee, 2010).

The research described in this thesis investigates how Signal Detection Theory and Thurstonian modelling can improve the effectiveness of sensory research in the FMCG industry, by exploring one specific test type of sensory tests; difference tests. Sensory difference tests are used to measure small differences between products, and can be used to answer important questions like: "*Are these two products similar in taste?*", "*Does this new ingredient make the product different?*", "*Will our consumers be able to notice the differences?*" The thesis focuses on two applications of Signal Detection Theory and Thurstonian Modelling. The first is to compare test methods, as there are many sensory difference tests available that largely differ in performance, and identify how to optimise the methods. With this knowledge, tailor-made methods can be designed for the specific test objectives and product types; tests that are more accurate and less costly. The second application of Signal Detection Theory and Thurstonian Modelling is to integrate results from different studies. This can improve the effectiveness of sensory testing, for example by relating sensory differences detected by a trained panel "*In Lab*" to differences found by consumers "*In Home*". Such knowledge can make future studies more predictive of what really matters to consumers, and improve the quality of decision making based on sensory results and reduce the overall number of tests that is required.

1.3 Overview of the thesis

The remainder of this thesis is organized as follows. Chapter 2 is a review of the current use of Signal Detection Theory and Thurstonian modelling for sensory science, and how it can be of benefit for the FMCG industry. It describes how different tests methods can be studied and optimised and requirements for what makes a good method for a specific objective. A case study with real data illustrates how results from different studies can be integrated to learn more about what sensory differences are important, so that better decisions can be based on the results. Chapter 2 is based on Van Hout, Dijksterhuis, and Groenen (submitted for publication).

In Chapter 3, we explore for the same-different test whether subjects can learn to use an optimal decision strategy. The same-different test is an alternative to the triangle and duo-trio tests and can have considerably more power if subjects use the optimal decision strategy. The main hypothesis in this chapter is that introduction of a 2-AFC-warm-up task, which is known to elicit an optimal decision strategy, would result in subjects continuing to use this strategy in the subsequent same-different test. The research was conducted with trained subjects using models systems and confirmed for some subjects that the optimal strategy in 2-AFC was carried over to the subsequent same-different test. Chapter 3 is based on Lee, Van Hout, Hautus, and O'Mahony (2007).

Using margarines, Chapter 4 compares three difference test methods for testing multiple products versus one reference product. The methods A-Not A with familiarization with the reference and voluntary reminder of reference during the test, A-Not A with prior familiarization with all products without reminder of the reference during the test, and similarity ranking of all products compared to the reference, have been performed by the same subjects. The results are compared in terms of R-indices. The test yielding the largest R-indices is the most sensitive in detecting differences. The differences in test performance can be explained in terms of the nature of the test, boundary variance, concept formation and cognitive strategies that are used. Chapter 4 is based on Lee, Van Hout and O'Mahony (2006).

Chapter 5 investigates the performance of three discrimination tests (A-Not A, 2-AFC and same-different) when discriminating between two different margarines. The effects of prior familiarization are investigated together with the effects giving a reminder of the reference product during the trials. Differences between test protocols were explained in terms of concept formation of the products, carry-over and fatigue effects, and memory problems caused by the time intervals between tastings. Chapter 5 is based on Lee, Van Hout and Hautus (2007).

Chapter 6 investigates more flexible methods for measuring overall sensory differences. The performance of three difference tests methods (A-Not A, 2-AFC and 2-AFCR) was compared in terms of learning effects over repeated sessions, again using margarines. This will determine how much training is needed before the subjects can perform the test consistently and emphasizes the importance of having effective familiarization procedures to stabilize test performance. Chapter 6 is based on Van Hout, Hautus and Lee (2011).

In the general discussion of Chapter 7 we will discuss the key findings and provide recommendations for further research.

2. FROM SENSORY PANELS TO CONSUMERS: UTILIZING SIGNAL DETECTION THEORY TO INCREASE THE EFFECTIVENESS OF SENSORY EVALUATION

Danielle H.A. van Hout, Garmt B. Dijksterhuis and Patrick J.F. Groenen

Submitted for publication.

Abstract

The ultimate goal of in-lab experiments is to predict consumer responses to products. A sensory panel is the closest in-lab approximation to consumers, but with conventional techniques the results of a trained sensory panel cannot be directly generalized to a market of consumers. In this paper, we present a framework for translating sensory panel findings to consumer perceptions. The framework makes use of the well known theories of signal detection and Thurstonian modelling that use d´ as a standardized measure between the measurements of two perceived intensities.

This review paper focuses on the most popular sensory difference tests. These tests are typically applied in situations where a product underwent a change in recipe, processing, or ingredients. The basic research problem is "Will the change to the product result in a sensory difference that will be perceived by consumers?" Often a trained sensory panel is used to test the effect of these differences. It is not automatically true that a perceivable difference between A and B discovered in the lab with a trained panel is also found in a consumer population.

With signal detection theory and Thurstonian modelling, data from different test methods and studies with different subjects can be compared using d´. Each test method implies psychological and physiological processes within subjects. With this information, methods can be assessed so that the most effective test for a specific purpose can be selected. In addition, d´ can also be used to compare groups of people, thereby allowing for a comparison of sensory panel findings with results of consumer tests. A case study illustrates how such a study can be designed in a cost effective way.

2.1 Introduction

Over the last decades, the role of sensory evaluation has become increasingly important for the fast moving consumer goods industry. Many business decisions are based on sensory test results, for example in quality control where sensory action standards specify to what extent batches of products are allowed to be different from a 'standard'. Such business decisions should be based on reliable test results that are predictive of consumers' natural reactions to the products. But what degree of deviation from the standard will consumers notice? Sensory difference tests are the powerful tools for building understanding of consumers' sensitivity to product differences. A problem is that some popular difference test methods, like the triangle test, are known to have severe issues, such as a significant lack of statistical power and robustness. These problems can lead to bad business decisions which in turn will result in loss of consumer loyalty, missed technical opportunities and decreasing sales. Fundamental research in the area of sensory difference testing will provide insights to understand current methods better and develop new and more effective methods, which in turn will further improve sensory evaluation so that it can be conducted more cost effectively with robust and predictive results, so that better decisions can be made based on sensory test results.

Sensory evaluation is conducted in various ways. Some companies have in-house trained sensory panels and/or untrained consumer panels, whilst others work with external agencies to outsource their research, or do both. O'Mahony (1995a) distinguishes between two classes of sensory evaluation with different goals, a classification that we will follow throughout this paper: Sensory Evaluation I (SE I), which is 'analytical' sensory evaluation, and Sensory Evaluation II (SE II), being 'consumer' sensory evaluation (O'Mahony, 1995a; O'Mahony and Rousseau, 2002).

- SE I refers to analytical measures of products' sensory properties. These tests are typically conducted with trained panels regarded to be human measurement instruments. Such panels often consist of relatively small numbers (8-16) of sensitive, well-trained subjects that only need a few repetitions to produce accurate results. SE I studies mostly take place under strictly controlled conditions to reduce noise as much as possible.
- SE II is used to study how consumers perceive the products, and should be conducted with subjects that are representative of the target consumers. Consumer panels show more variance than trained sensory panels. The number of subjects required for consumer tests is relatively large (typically 50 to 100) and needs to be increased when subgroups need to be compared (e.g. users versus non-users). Replicating the measurements with consumers should be handled with care. Too many repetitions will make the consumers more sensitive, and thereby they will fail to represent the target consumer group. SE II tests should also be 'ecologically valid' so they should

preferably take place in conditions similar to the natural conditions of usage. In general, for SE II it is important to find the right balance between controlling the test conditions and an ecologically valid test setting.

Both types of sensory evaluation are important and serve different needs. SE I is needed to measure differences between products and gain understanding of how changes in product formulation or processing influence sensory properties. However, with only SE I, there is no understanding of the consumer relevance of product differences. SE II is required to measure if and how target consumers perceive the product changes, but using only SE II would be very expensive, as there are large sample sizes of subjects required and results will be target-group (e.g. country, user type) specific and therefore require considerably more testing. Sensory difference tests can be used for both SE I and II, but both types have different requirements for a test to be effective. Therefore it will not be easy to identify a difference test method suitable for each purpose. Instead of such a 'one-tool-fits-all' approach, it is more advantageous to have tailor-made analysis tools that allow us to integrate results from different types of test methods.

The biggest impact on the development of sensory difference test methods was made using psychophysical theories. Signal detection theory and Thurstonian modelling made it possible to study human performance in sensory tests in a systematic way and provide an underlying theoretical framework for sensory and consumer test methods (O'Mahony, 1992). The basic measure in psychophysics, *d´* (*d*-prime), is used to express sizes of sensory differences between products and makes it possible to accurately quantify small sensory differences. *d´* can be seen as a generalised measure of sensory difference (effect size) that is independent of the test method used. Therefore, *d´* can be used to accurately and systematically compare sensory tests and study the effects of changes in test design and instructions on the performance of the test. In this manner, sensory difference tests can be optimised and made more efficient, so that fewer tests are required for obtaining results of similar quality (Ennis, 1993; Kim, Jeon, Kim and O'Mahony, 2006a; Lee and O'Mahony, 2006, 2007). For companies there are many benefits to use approaches based on signal detection theory and Thurstonian modelling for their sensory evaluation practises. By doing so, sensory evaluation will become cheaper and more accurate. Next, these models also allow companies to expand their strategic knowledge of products and consumers, as it becomes easier to gain systematic understanding of how product differences influence consumers' perceptions. Unfortunately, outside academic research, these models have often been regarded as too complex and have remained underused in industrial sensory evaluation.

Several papers have been published that review applications of signal detection theory and Thurstonian modelling. These papers vary from very practical to highly abstract. For example, O'Mahony and Rousseau (2002) provide a practical overview of sensory discrimination tests and explain how Thurstonian approaches can be applied to study various experimental factors important to selecting the most suitable sensory different tests. They recommend which methods to use for what purpose, and give specific instructions on how to calculate *d´*-values from test results. Lee and O' Mahony (2004) give an introduction to the ideas and applications of Thurstonian modelling as applied to sensory difference testing, and explain how memory effects and in-mouth processes can influence *d´* values. At a more abstract level, Ennis (1998) describes how signal detection theory and Thurstonian modelling can be used as a foundation for sensory and consumer science.

The aim of this review is to address how signal detection theory and Thurstonian modelling can be utilised to answer specific business questions and decisions. The two important questions that will be discussed are 'How does one select the best sensory difference tests for a specific purpose?' and 'What sizes of product differences are consumers able to detect?' and although the answers to these questions are very situation and product specific, this paper provides some practical guidance and recommendations based on insights from sensory science and psychophysics. Factors that make sensory difference tests effective for SE I and II will be discussed, and the strengths and weaknesses of various difference tests reviewed.

The chapter is organized in the following way. Section 2.2 provides a concise explanation of the elements of signal detection theory and Thurstonian modelling that are important for sensory difference tests. Section 2.3 describes how to optimally implement difference tests for use in SE I and II. Section 2.4 presents a case study that compares the sensitivity of consumers (SE II) to the sensitivity of an in-house trained sensory panel (SE I). Section 2.5 contains a discussion and conclusions.

2.2 Signal detection theory and Thurstonian modelling

To be able to understand the benefits of signal detection theory and Thurstonian modeling, we provide a concise explanation of the basic elements of this field that are important for sensory difference tests. For a detailed overview of the underlying principles, we refer the reader to the book by Macmillan and Creelman (2005). This section focuses on how these models differ from conventional statistics and how they can be used to analyse, understand, improve, and interpret sensory difference test methods and results.

Conventional approaches for sensory data analysis are based on *deterministic* statistics, also called *chance* or *guessing* models, which consider people's sensory perception of products to be fixed points. A subject giving a wrong answer in a difference test, could not detect a difference and was forced to give an answer, had to guess, and guessed wrong (Rousseau, 2001). One problem of using deterministic models for sensory data is that they are unable to explain why people perform certain tests better than others. For example, in both the 3-AFC and in the triangle test, three samples are being tested (either two of product A and one of product B, or one of product A and two of product B). The instruction in the triangle test is: 'Which is different from the other two?' The instruction in the 3-AFC is: 'Which one is more/less intense (of 'something', e.g. sweet, brown, or fruity, etc.) than the other two?' The chance of guessing correctly is 1/3 for both tests. However, when comparing the performance in studies with the same set of products and people, the 3-AFC consistently yields more correct responses. This does not mean that the sensory difference between the products is larger in 3-AFC. It only means that 3-AFC is more powerful and that it is more likely that significant differences are being obtained by 3-AFC than by the triangle test (Frijters, 1979a, 1979b; O'Mahony, Masuoka and Ishii, 1994; Ennis, 1998). Statistical significance alone is not a good criterion to base conclusions on. It is a measure of strength of the evidence against the null hypothesis of no difference. Therefore statistical significance is not the same as the size of the difference. Simply enlarging the sample size increases statistical significance but does not change the true size of the difference. In deterministic statistics the 'Proportion of correct responses' (P_c) is used to represent the size of differences. The problem of P_c is that it varies by test type, as explained with the example of the 3-AFC and triangle test. This variation makes it impossible to compare results in P_c between different test types. Another problem is that P_c is strongly affected by response bias, which is present in every test and is caused by peoples' tendency to use some responses more often than others. Response bias creates additional noise in the P_c measure and thereby reduces the accuracy of the results (Macmillan and Creelman, 2005).

Signal detection theory and Thurstonian modelling are *probabilistic* models. The foundation of Thurstonian modelling was laid by a model for perceptual measurements, published as the ‘Law of Comparative Judgements’ (Thurstone, 1927), which proposes that when a subject experiences a stimulus, for example, when listening to a sound or tasting a product, the perceived intensity varies intrinsically. Therefore, the perceived intensity of a product is not considered fixed but variable, and is regarded as a draw from a probability (normal) distribution, hence the name *probabilistic* models. Thurstonian modelling is closely related to signal detection theory, which is derived from the field of electrical (communication) engineering. Signal detection theory treats the human nervous system as a communication system (Green and Swets, 1966) and determines how well an individual can detect a signal embedded in noise in a sensory task, while taking into account two factors; the *perceptual variance inherent in the sensory system*, and the *effectiveness of the decision strategies* that people use when performing the task. In sensory science the terms signal detection theory and Thurstonian modelling are often used interchangeably, and together they provide an integrated framework for studying and understanding the mechanisms of sensory measurements (O’Mahony and Rousseau, 2002; Lee and O’Mahony, 2004). These models have been popular in the field of visual and auditory psychophysics (Hautus, Irwin and Sutherland, 1994) as well as in sensory science (O’Mahony, 1992; Ennis, 1993).

The first factor in signal detection models is the *perceptual variance*. A product’s intensity is sometimes perceived as weaker and sometimes as stronger. This variation is caused by several internal and external factors, such as the strengths of the signals in the brain, the inherent product variation, a person’s mood, and factors related to memory and sensory adaptation. These are reasons to view the perceived intensity not as a fixed point but as varying like if drawn from a normal distribution (see Figure 2.1).

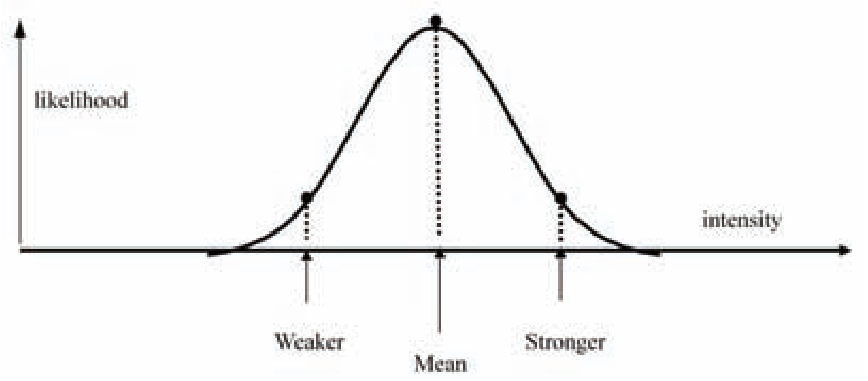

Figure 2.1 The intensity of the sensory perception (along the x-axis) of a stimulus varies according to a normal distribution. The y-axis contains a measure of the probability of obtaining a certain sensory intensity.

The perceptual variance is used to create the scale units for expressing differences between products (Figure 2.2), expressed in units of standard deviation (*s*) between the means of the two distributions (m_1, m_2). Defined in Formula (1) (Macmillan and Creelman, 2005):

$$\boldsymbol{d' = \frac{m_1 - m_2}{s}} \tag{1}$$

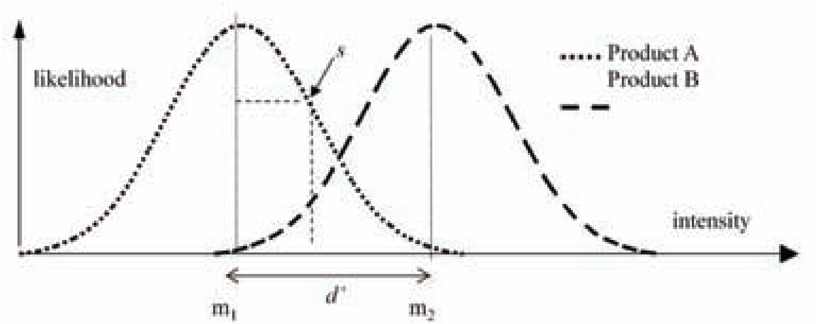

Figure 2.2 The intensity distributions of two confusable products. The distance between the two means, in units of *s* is 2.5. This is the measure *d´*, the universal size of sensory difference.

The measure *d´* is a generalized measure of size of sensory differences that is independent of test method that is used. This makes it possible to directly compare test results from different tests. Another major advantage is that *d´* is unaffected by response bias in contrast to conventional measures like P_c. As *d´* is untainted by response bias, it can be used more accurately for comparing sensitivities of (groups of) consumers or of different test methods (Macmillan and Creelman, 2005). In psychophysics, *d´* can be calculated as a measure of sensitivity of each individual subject or for a group of subjects, in sensory and consumer tests *d´* values are often used as sensitivity measures for the total sensory panel or for specific segments of consumers. With signal detection models and the use of *d´*, it is possible to study test designs and experimental variables in terms of their effect on test performance (O'Mahony and Rousseau, 2002; Rousseau and O'Mahony, 2000). When the variation is larger (larger *s*) the difference that can be perceived will be smaller (smaller *d'*). This means we can hypothesize how test design factors will influence test performance. For example, if there are more samples included in a test, there will be more fatigue and carry-over effects. These effects introduce more variance thereby reducing the test performance.

The second factor included in signal detection models is the *cognitive decision strategy* used by a subject performing the test. The decision strategy is based on the instructions in the task, the presentation of the products and the subjects' knowledge of the sensory evidence. Certain strategies are more effective than others and lead to a higher P_c. If the decision strategy that subjects use in the test is known, it can be accounted for and accurate *d´* values can be calculated for the test. Unknown or unstable decision strategies that vary between people or change after repeating the test, can lead to inaccurate estimates of *d´* (O'Mahony 1995b).

There are two classes of decision strategies that are frequently used in sensory difference tests. The first class of decision strategies consists of *beta* or *identification* strategies. In psychology this class of strategies is referred to as the independent observation model (Macmillan and Creelman, 2005). When subjects use a beta strategy they evaluate the product independently and make a decision. The beta strategy is typically used in an A-Not A test. In this test, subjects receive one product at a time with the instructions "is this product A or not A?" The strategy will typically be to identify each product presented as product A or not A, based on a perceptual boundary. If the intensity is lower than the boundary criterion, subject will categorize it as product A, if the intensity is higher than the boundary criterion, the subject will categorize the product as not A (see Figure 2.3).

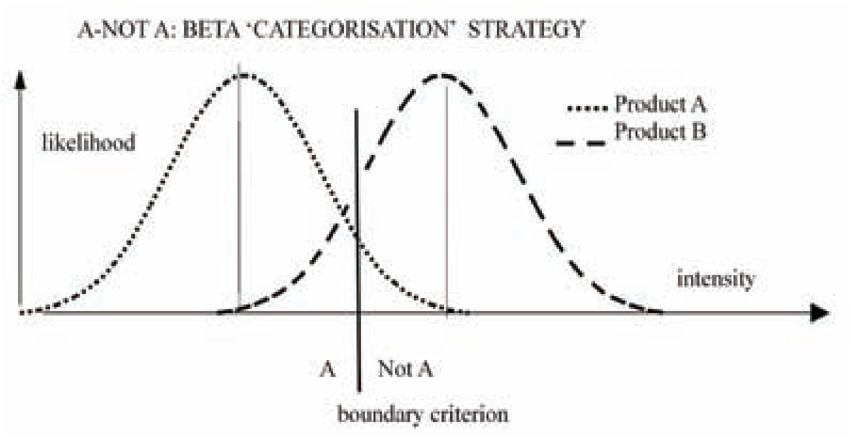

Figure 2.3 Beta - Identification strategy in the A-Not A test: if the perceived intensity of a sample is lower than the boundary criterion, the response will be "A"; if the intensity is higher than the boundary criterion, the response will be "Not A".

The second class of strategies are the *tau strategies*, also called *differencing* strategies (Frijters, Kooistra and Vereijken, 1980; O'Mahony and Hautus, 2008). When a subject uses a tau strategy, a comparison is made between the products to make a decision. For example, in a 3-AFC test, subjects receive three products with instructions such as "One product is sweet, the others are not. Which is the sweet one?" In this test, typically a skimming strategy is used. People skim off the product with the highest sweetness (Figure 2.4).

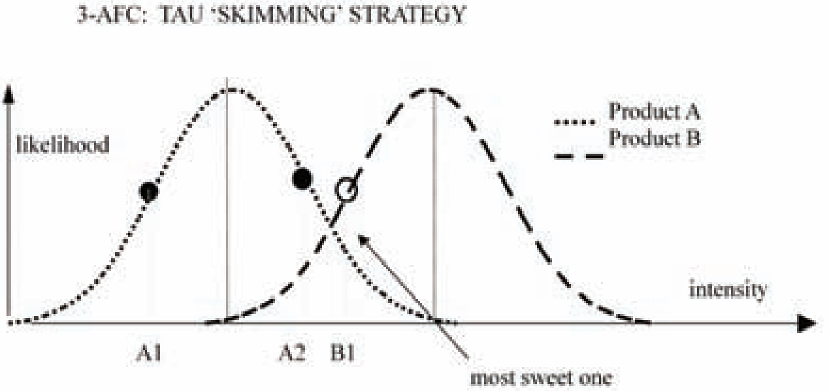

Figure 2.4 Tau - Skimming strategy in the 3-AFC test: perceived intensities of 3 samples (shown as circles on the distribution) in this situation the subject would correctly skim of product B (the open circle) as the sweeter one.

The skimming strategy is a more optimal tau strategy than another tau strategy which is mostly used in the triangle test, the so-called 'Comparison of Distances' (COD) strategy. The instructions in the triangle test are: "Two are the same, one is different. Which is the different one?" subjects use the tau strategy to determine the differences between the products and identify the one that is most different from the other two (see Figure 2.5).

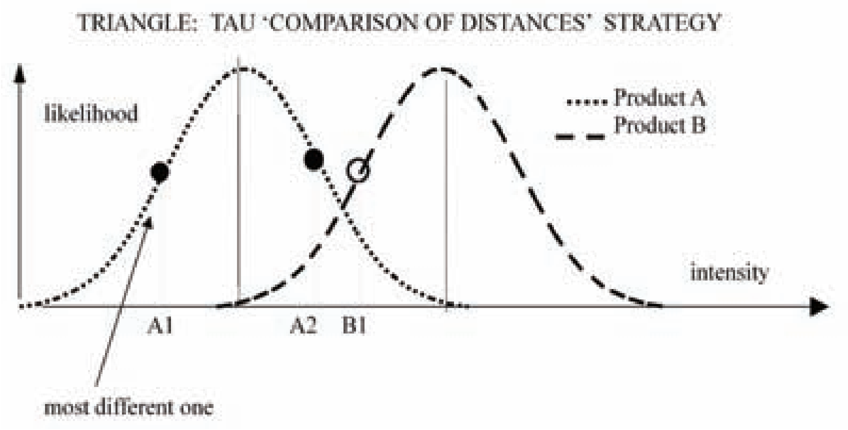

Figure 2.5 Tau - Comparison of Distances (COD) strategy in the triangle test: perceived intensities of 3 samples, in this the subject would identify the product A1 as the most different one because the difference to both other products is the largest. The response would be incorrect.

Optimal decision strategies lead to better test performance, by allowing subjects to use an easier way of cognitive processing. In studies comparing the two strategies in the 3-AFC and triangle tests with the same products, the 3-AFC test consistently resulted in a larger P_c than the triangle test (Frijters, 1979a, 1979b, 1982; O'Mahony, 1995b). Table 2.1 shows the proportion of correct responses for four test methods for a variety of *d´* values.

Table 2.1 The proportion of correct responses for the triangle, 3-AFC, duo-trio, and 2-AFC tests for a variety of *d´* values (Ennis 1993)

d´	*P_c (proportion of correct responses)*			
	Triangle	3-AFC	Duo-Trio	2-AFC
0.0	0.33	0.33	0.50	0.50
0.5	0.35	0.48	0.52	0.64
1.0	0.42	0.63	0.58	0.76
1.5	0.51	0.77	0.66	0.86
2.0	0.60	0.87	0.75	0.92
2.5	0.70	0.93	0.82	0.96
3.0	0.78	0.97	0.88	0.98

In psychophysics, *d´* = 1 is often used as an empirical threshold, or just noticeable difference (JND) and compares to P_c = 76% in a 2-AFC test (O'Mahony and Rousseau, 2002; Green and Swets, 1966; Macmillan and Creelman, 2005). In sensory science, such a *d´* value can also be used as an initial action standard and may be updated once more accurate knowledge has been established on the effects of product differences on consumers' perception.

Variation in the use of decision strategies in a test has a direct effect of lowering the robustness of the test. Therefore, in order to measure differences accurately, a stable use of decision strategies in the test is essential. For many of the common difference test methods, signal detection models for calculating *d´* were developed and published (Ennis, 1990, 1993; Ennis and Jesionka 2011; Bi and Ennis, 2001a, 2001b; Hautus, Sheperd and Peng, 2011a, 2011b; Hautus, Van Hout and Lee, 2009). The use of the signal detection measure *d´* makes it possible to directly compare results across different test protocols and helps to make more accurate interpretations of the impact of certain changes in test design. In such studies, the new or adapted test method is typically compared with an existing method of which the signal detection model is known, using the same products, and with the same subjects performing both tests in balanced order, *d´* values can be calculated and compared. If the *d´* values are significantly different, this means that the signal detection model of the new test is not accurate enough and not all cognitive and physiological factors have been taken account for in the model.

2.3 Optimal implementation of difference tests, a signal detection approach

Sensory tests should be statistically powerful, so that only a low number of trials are required. Also, cognitive and physiological factors that can potentially lower the tests' *operational power* (Kim, Chae, Van Hout, and Lee, 2011; Van Hout et al. 2011 (chapter 6)), should be considered and optimised so that results are *robust,* meaning that if the test would be repeated, similar conclusions would be drawn. Further requirements of what makes a test effective depends on the goal; is it SE I or II?

2.3.1 Designing difference tests

Sensory difference tests consist of three stages: (1) the *introduction* stage, (2) the *task and response* stage, and (3) the *feedback* stage. Careful design of each of these stages provides opportunities to optimize the tests' effectiveness for specific purposes.

The *introduction* stage takes place prior to the actual testing and defines the test context for the subjects. In the shortest introduction, subjects only get instructions on how to perform the task and evaluate the samples. A lengthier introduction could consist of a familiarization with one or more test samples. For SE I, it is usually advantageous to perform a familiarization procedure. Proper familiarisation with the samples during the introduction will improve and stabilize the panel's sensitivity and increase the test's operational power. Such a familiarization can be conducted with only one product, the 'reference', or with more products. Typically, a familiarization with both the reference and the test products will allow for faster learning (Lee, Van Hout and O'Mahony, 2006 (chapter 4); Lee, Van Hout and Hautus, 2007b (chapter 5)) depending on the product type; for example products with intense flavour can desensitize the subjects (Rousseau, Rogeaux and O'Mahony, 1999). Lee, Van Hout, Hautus, and O'Mahony (2007a, Chapter 3) found that after a warm-up task consisting of a 2-AFC test, the optimal decision strategy use in 2-AFC was carried over to the subsequent same-different test, making the test more operationally powerful. They hypothesised that the positive effect of such a warm-up task to the same-different test, may also apply to the A-Not A test. In further research with the A-Not A test, Kim, et al. (2011) found that when a trained sensory panel is familiarized with a reference product labelled with a (fake) brand image, their performance increased faster than without a brand image. As SE II aims at reliably predicting consumers' natural perception of product differences, training is not advisable because it would make the sample of consumers more sensitive than they would be naturally, rendering them unrepresentative of the consumer population. However, exposing consumers to the products during the introduction will give them an idea of what differences can be expected, so that they can adjust their criterion (Rousseau, Stroh and O’Mahony, 2002). Also, when a study is conducted with (loyal) users of the product, an example of the familiar product during the introduction will refresh their memory. Kim, Chae, Van Hout and Lee (2014) found that when consumers are somehow more affectively involved in the task, by using a brand logo or a preference task during the introduction, their sensitivity increased. This is in line with previous findings of better performance in a so-called ‘authenticity test’ in which subjects perform the evaluation affectively, instead of the traditional more analytical evaluation (Wolf Frandsen, Dijksterhuis, Brockhoff, Holm, Nielsen, and Martens, 2003; Boutrolle, Delarue, Köster, Aranz, and Danzart, 2009; Chae, Lee, and Lee, 2010).

The middle stage is the *task and response* stage. This stage defines the test proper. Each test method requires specific instructions to be given to the subjects. Some tests use a reference product and others not, and in some tests a reminder of the reference is available during the trial for direct comparison with the test sample. Methods differ in the number of samples and responses generated by a trial. For example, in an A-Not A test each sample generates one response, in a 2-AFC test two samples generate one response, and in a triangle test three samples generate one response. The defining characteristics of seven popular sensory difference tests are summarised in Table 2.2.

Table 2.2 Task and response structures of seven popular test methods

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

Irrespective of the test method used, samples in a trial can be given to the subject simultaneously, or one after another in the same session, or one each day or week. This mainly depends on the product type. When samples are presented in a trial spread over a longer time, the task will be more demanding of memory. Delayed presentation of the second sample decreases performance in the same-different test, especially when subjects are unfamiliar with the samples (Cubero, Avancini, and O'Mahony 1995; Avancini, Cubero, and O'Mahony, 1999). Four of the tests, 2-AFC, 2-AFCR, triangle and duo-trio, are forced choice methods that require subjects to make a choice between samples. In this way, decision criteria are stabilized between samples and this reduces response bias. The other three tests, A-Not A, A-Not AR and the same-different test, require a categorical response and leave subjects free in setting their criterion, for responding if a product is A or not A, or if two products are the "same" or "different". In such tests typically a sureness rating is added for subjects to respond how sure they are of their answer (O'Mahony, 1995; Lee and Van Hout, 2009). This makes it possible to calculate sensitivity (in *d´*) independent of response bias, stabilizes response bias, and lowers the risk of floor and ceiling effects. For forced choice tests a sureness rating is not required but these ratings can be added to every test as a diagnostic tool to explore effects of changes in design or instructions (O'Mahony and Hautus 2008; Wichchukit and O'Mahony 2010; Santosa, Hautus and O'Mahony 2011).

The final stage of a difference test is the *feedback* stage. It is not always considered to be part of the test, but feedback can play an important role. In visual or auditory psychophysics, feedback on whether the response was correct or incorrect is often given to subjects and increases performance in subsequent tests. For SE I, feedback can help the subjects in the sensory panel to perform better faster and should be considered as an integral part of difference test experiments as it can help to stabilize performance over repetitions. In SE II, feedback is usually not advised as it would increase consumers' performance and therefore render the test less predictive of reality.

2.3.2 Test effectiveness

The two factors earlier mentioned as important for a test's effectiveness, the operational power and the robustness of the results, are influenced by various physiological and cognitive sources of variation. One important and undesired source of variance that reduces the tests sensitivity are *sequence effects*. Table 2.3 provides an overview of the number and type of sequences of the 7 test methods, already described in Table 2.2. Estimation is given of the relative strength of the sequence effects in the different methods.

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

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

Each test method has a number of sequences in which the products are evaluated. Some sequences perform better than others, a phenomenon which was repeatedly found and analysed using sequential sensitivity analysis (O'Mahony and Goldstein, 1996, Rousseau and O'Mahony 2001, Santosa and O'Mahony, 2008; Lee and O'Mahony 2007a). The more products and more sequences, the stronger the sequence effects will be. The difference in performance of different sequences can be quite large, for example, Kim et al. (2013) found differences in results ranging from d'=1.21 (σ=0.18) to d'=2.04 (σ=0.15) between the best and worse sequence in triangle tests, the 2-AFCR did not suffer from such sequence effects. The more samples or more sequences, the larger the variance induced by physiological and cognitive factors such as sensory fatigue, carry over, memory and adaptation. Therefore a test with a low amount of sequences, like A-Not A or 2-AFC will be less affected by sequence effects than tests with more products and sequences like same-different, duo-trio, and triangle. Products that are more intense in nature, like for example mustard, are more affected by these effects (Rousseau, et al. 1999)

Another factor that influences test performance is the *decision strategy* used. When the decision strategy used in a test is known, a signal detection model can be constructed to calculate *d'* from the responses. For a particular test, some decision strategies are more powerful than others and require fewer trials to detect a difference. Also the stability of the strategy is important. Strategy shifts add variance and lower test sensitivity and robustness of results. Suppose all subjects start using a certain decision strategy, but if some of them change to a more effective strategy when they get more experience with the products or task, differences will start to appear larger. For example, in a triangle test with 40 subjects, 42% correct responses, would be a *d´*=1.7 when a COD strategy is used, but only a *d´*=0.8 if subjects would use a skimming strategy. Recent research in this field has revealed clear insights on the type and stability of the decision strategies used in test methods. Table 2.4 shows for the seven tests the most likely decision strategies.

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

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

The A-Not A test typically evokes a beta strategy. Provided that subjects are familiar with the reference they can categorize a product by deciding whether it is the same product as the reference or not. Also in the A-Not AR test, typically a beta strategy is used. However, when subjects are insufficiently familiarised with the reference, some subjects use a tau strategy for A-Not AR, thereby lowering test performance (Hautus, Shepherd and Peng, 2011). Stocks, Van Hout and Hautus (2013, 2014) found with A-Not AR, that simple stimuli, like taste model solutions, evoked both beta and tau strategies by trained panellists, in a 60%/40% ratio. However, with more complex stimuli, such as food and beverages, the trained panellists consistently used beta strategies. This indicates that more realistic stimuli allow subjects to make more optimal decisions. The same-different test normally evokes a tau strategy. Santosa, Hautus and O'Mahony (2011) demonstrated that this use of the tau strategy was consistent over judges and stable after repetitions. For the 2-AFC and 2-AFCR tests, both beta and tau strategies are equally powerful and lead to the same results in *d´* (Hautus et al. 2009, 2011a, 2011b). In the triangle and duo trio test, typically the tau COD strategy is used, but strategy shifts are frequently reported (O'Mahony, 1995b). Re-tasting samples in a trial or performing repetitions will allow some subjects to use more optimal strategies (Rousseau, 2001). The signal detection model for triangle tests, assuming a COD strategy, is then not valid anymore and gives incorrect, overestimated, *d´* values. In terms of decision strategies used, the A-Not A, 2-AFC and 2-AFCR tests evoke the most optimal and stable strategies, the same-different test evokes sub-optimal but stable strategies, whereas the triangle and duo-trio tests evoke sub-optimal and less stable strategies.

Each test method has an optimal performance level, also called *operational window*, indicating the range of differences that can be measured most accurately (Hautus and Lee, 1998; Jesteadt, 2005). Table 2.5 gives an overview of the optimal performance level of each test and the required sample sizes for detecting differences.

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

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

Some tests, like the 2-AFC, are better at detecting small differences but are prone to *ceiling effects* which make larger *d´* less accurate (Macmillan and Creelman, 2005; Van Hout et al., 2011). Others, like the same-different test, can accurately measure relatively large differences, but have *floor effects*, which means that they are not sensitive enough to pick up small differences consistently.

2.3.3 Selecting the best test for SE I and II

The triangle and duo-trio test have low operational power and are therefore only suitable for detecting larger sensory differences (*d´* =2). These tests are also prone to suffer from cognitive and physiological sources of variation, like sequence effects and decision strategy shifts, that hamper the robustness of results (O'Mahony, 1995b; Kim et al., 2006, 2012). The same-different test would be a better alternative, as this test evokes consistent use of the tau strategy over repeated trials, and can be used both for SE I and II. For detecting smaller differences (*d´*=0.5-1) the A-Not A and 2-AFC tests, as well as their reminder versions, A-Not AR and 2-AFCR are generally more effective. These tests require a lower number of trials to provide robust results (Santosa et al., 2011; Kim et al. 2011). Using a reminder has advantages and disadvantages as there are, for example, fewer loads on memory, but sequence effects are larger due to the extra product presented. Whether using a reminder is effective depends on many factors, such as the product type and test objectives, and can be determined experimentally for new product types, so that the most effective test protocol can be used in further testing.

In SE I, as subjects learn about the task and products, performance will increase to a maximum level (asymptotic performance). Familiarization with the products during the introduction stage can help to reach this asymptotic performance as early in the experiment as possible (Hautus et al., 2011a, 2011b). Feedback to subjects after each trial can also increase the speed of learning (Stocks et al., 2013, 2014). The 2-AFCR test demonstrated stable performance in early trials, while performance in A-Not A took more time to increase (Van Hout et al., 2011). This finding suggests that in cases of few trials, with many changes of reference product, the 2-AFCR test would be better to use. The A- Not A test is particularly economical if the same reference is used in more trials or when multiple products need to be compared to one reference product (Lee, Van Hout, Hautus, and O'Mahony, 2007a (Chapter 3)).

In SE II, it is important to select subjects that have a similar sensitivity as the target group of consumers. Factors that influence performance, such as extensive training, should be avoided. Task and test instructions need to be easy and straightforward as we want to test whether consumers perceive product differences, not if they can perform complex cognitive tasks. The 2-AFCR and the A-Not A tests are recommended for detecting small differences, the latter especially when testing a well-known product with consumers. An introduction with the reference product in the test setting is required for the A-Not A test, but not necessary for 2-AFCR test as in 2-AFCR each trial starts with the reference product. The monadic assessment of the A-Not A test reflects reality as consumers normally experience the products one at a time and compare it to their memory of previous ('reference') products, they usually do not taste two products side-by-side and compare them directly. When products are unfamiliar, the 2-AFCR is a safer choice than A-Not A, as it depends less on the memory representation of the reference.

2.4 From panel to consumers: A case study relating SE I and II

2.4.1 Introduction

When is a difference between products important enough to warrant action? This is probably one of the most important questions to address, but also one of the most difficult. A case study, based on empirical data, illustrates how signal detection applications can help a company to optimize their action standards to improve the quality of the business decisions based on sensory research. For a series of margarines, differences between products as perceived by consumers (SE II) are related to the differences measured in a trained sensory panel (SE I). Understanding this relationship makes it possible to use the sensory panel for screening products and predicting whether consumers will be able to perceive the differences (Isshi, Kawaguchi, O'Mahony and Rousseau, 2007). Objectives of this case study are to investigate what difference between products is perceivable by consumers and to model the relationship between the sensitivities of a trained panel and of consumers. This information can be used to define consumer relevant action standards for the trained sensory panel.

2.4.2 Materials and methods

The SE I experiments were performed by an in-house trained sensory panel that consisted of 12 subjects, screened for their sensory sensitivity and with at least two years of weekly experience with performing various sensory difference and descriptive tests. For SE II the experiments were conducted with a random sample of 100 consumers for each experiment, recruited from two different regions in Germany. In the experiments, the following products were included:

- In Experiment 1, seven commercial margarines from the European market were used with small variations in texture as well as in flavour. These products will be referred to as A, B, C, D, E, F, and G.
- In Experiment 2, two sets or products were used. Set 2a consisted of two commercial products from Experiment 1 (A and D), and two mixtures of the two products. These products will be referred to as A, D, Ad (mixture of two thirds of A and one third of D), and Da (mixture of one third of A and two thirds of D). Set 2b consisted of one of the commercial products (F) selected from Experiment 1 and with two different levels of salt added. Products will be referred to as F, F1 and F2.

The consumer sample was stratified for gender and age. The A-Not AR test method was used for all experiments. In the introduction stage, subjects were familiarized with the reference product, by tasting it four times in a row. The samples, each with a reminder sample of the reference, were given to the subjects in a randomised order. Subjects needed to indicate whether the product was the reference or not, and how sure they were, on a 6 point scale with the categories: "reference, sure", "reference, not sure", "do not know but guess it is the reference", "do not know but guess it is not the reference", "not the reference, not sure", "not the reference, sure". The consumers each tested all products without repetitions; the trained sensory panellists tested all products eight times.

2.4.3 Data analysis

The responses of the consumers in the A-Not AR test were pooled and *d*’ estimates and standard deviations were calculated by fitting receiver operating characteristic (ROC) curves. This was accomplished using the software, SDT Assistant (Hautus, 2012), to fit the appropriate signal detection theory (SDT) model to the data using maximum-likelihood estimation. The likelihood function employed was that developed by Dorfman and Alf, 1969). For the sensory panel data, two approaches are possible for calculating *d´* for a group of subjects. The first method is to calculate *d´* for each individual subject and then average over individuals. The second method is to pool the data from all subjects and calculate a pooled *d´*. It depends on the variation in the group of subjects and the number of repetitions which of the two methods is optimal. Hautus (1997) found that when the number of repetitions is low, a pooled *d´* would be less biased. With many replications per subject, the average *d´* is less biased. The general recommendation is to calculate *d´* in both ways and use the highest, as it has the greatest likelihood of being least biased. In this study, pooling resulted of the highest *d´* values and would, therefore, be the most accurate estimate of the true *d´* (Hautus, 1997). Linear regression analyses were carried out for each product set to investigate the relationship between the SE I and II results, and to determine if relationships found were consistent.

2.4.4 Results

The *d´* values and standard deviations for the SE I and II studies are listed in Table 2.6. Different letters are used to indicate significant differences (*p*=0.05) between *d´* values.

Table 2.6: Sensory differences between the samples and the reference product, as measured by a sensory panel (SE I) and by consumers (SE II).

Experiment		Results SE I		Results SE II	
Set	Samples	*d*’	std	*d*’	std
1	Ref(A)-B	2.77a*	0.26	1.24a	0.16
1	Ref(A)-C	1.77b	0.17	0.53b	0.15
1	Ref(A)-D	1.81b	0.2	0.78ab	0.15
1	Ref(A)-E	2.62ab	0.26	0.77ab	0.15
1	Ref(A)-F	2.25ab	0.22	0.84ab	0.15
1	Ref(A)-G	2.17ab	0.22	1.08ab	0.16
2a	Ref(A)-D	1.79a	0.21	0.72a	0.16
2a	Ref(A)-Ad	0.70b	0.15	0.38a	0.15
2a	Ref(A)-Da	1.83a	0.21	0.53a	0.15
2b	Ref(F)-F1	0.63a	0.17	0.49a	0.16
2b	Ref(F)-F2	1.28a	0.21	0.74a	0.17

* Different letters are used to indicate significant differences (p=0.05) between *d´* values.

For all products, sensory differences detected by the sensory panel are larger than those detected by consumers. Differences between products A and D were measured in both experiments and results were very similar (1.81 and 1.79 in the sensory panel, and 0.78 and 0.72 for consumers).

A linear regression analysis was conducted with the two sets of *d´* values to predict the *d´* value of consumers from the *d´* value of a trained panel, Figure 2.5 shows the *d'* values of the product differences and regression line with 95% confidence interval, illustrating the relationship between the sensory panel and consumers. The correlation between the two measures is .70, indicating that the *d´* value of consumers can be reasonably predicted from the *d´* of a trained panel.

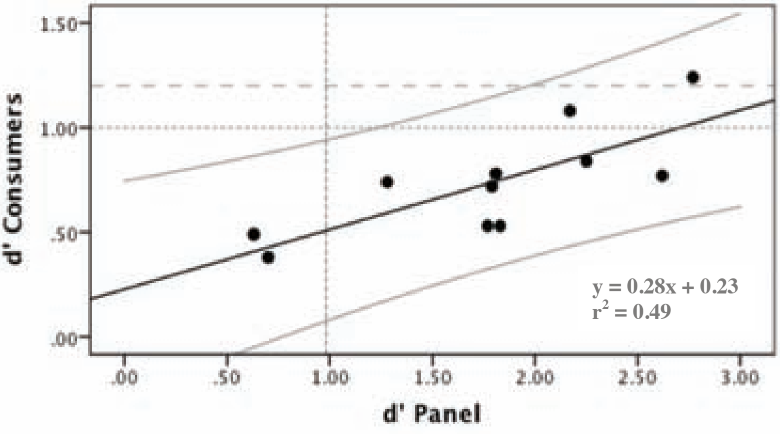

Figure 2.6 Scatter plot of *d´* values, regression line of trained panels predicting sensitivity of consumers, and the two curved lines indicating the 95% confidence prediction interval. The original SE I based action standard of *d´* = 1 is visualized by the vertical grey dotted line. The SE II consumer relevant action standards of *d´* = 1 and 1.2 are visualized by the horizontal grey dotted lines.

This study shows that the sensory panel is more sensitive than consumers as the regression coefficient is significantly smaller than 1 (95% confidence interval for the slope is 0.11-0.46). With this information, action standards can be optimised so that testing can be more effective. Assume the simplistic situation that the original action standards were soley based on SE I results and set at *d´* =1.

- For products to be regarded as *similar*, *d´* should be significantly lower than 1, meaning the 95% confidence prediction interval should not include 1. In practise this would mean that differences of *d´*<0.65 would be regarded as similar.

- For products to be regarded as *different*, *d´* and 95% confidence prediction interval should be equal to 1 or higher, which means in practise that when differences are *d´* ≥0.65 the products are regarded as different.

This study provided new information on consumer sensitivity to product differences, based on which it could be decided to define new action standards with values that are predictive of consumers' perception.

- For products to be *similar*, the consumer *d´* value should be lower than 1, taking into account the 95% confidence interval. Translated to differences in SE I, it can be seen that when the upper confidence interval reaches *d´* =1 for consumers, this difference in the sensory panel is *d´* =1.3. This means that for sensory differences with a *d´* larger than 1.3 in the sensory panel we cannot exclude that a *d´* =1 will be found with consumers and, thus, the difference could be noticeable by consumers. The new SE I action standard for products to be regarded as similar can therefore be changed to *d´* <1.3.
- For products to be regarded as *different*, one could decide that the difference should at least be *d´* =1.20 for consumers, as this means that over 80% of consumers would be able to detect the differences (*d´* = 1.20 compares to P_c = 0.8 in a 2-AFC test (Ennis, 1993)). Translated to differences in SE I, from the graph in Figure 2.5, it can be determined that the new SE I action standard for products to be different could be changed to the point where the upper line of the confidence interval reaches *d´* =1.2 for consumers, this is at *d´* =1.9. The new SE I action standard for products to be regarded as different can therefore be changed to *d'*≥1.9.

In Table 2.7 this example how action standards could be updated is summarized, and it shows how the change to the new action standards reduces the number of trials that are required (Bi and Ennis 2001a, 2001b).

Table 2.7: the old and new SE I action standards, and the number of trials required to achieve a power of 0.8, and α= 0.05

	"Similar"	Number of trials required	"Different"	Number of trials required
Old action standard	*d´*<0.65	57	*d´*≥0.65	57
New action standard	*d´*<1.3	20	*d´*≥1.9	<17

In contrast to the original action standards, which are soley based on sensory panel results, the new action standards are predictive of whether consumers would perceive the differences. The new action standards are also larger than the original action standards, resulting in a large reduction in the number of trials required.

2.4.5 Conclusions and recommendations

The case study illustrates how increased understanding of consumer sensitivities to product differences can be used to optimise action standards and make sensory research more time and cost effective. More samples will lead to better predictive models of the relationship of SE I and II. Also it is important to note that the relationship that was found between trained panels and consumers is only valid for this type of products and consumers. When using different products, or predicting other consumer groups, it is very likely that the relationship will be different.

Another finding from this study is that the future design of similar experiments can be improved. The variance in SE I results can be further reduced by better familiarization, for example providing a brand logo with the reference product, or by familiarizing with both reference and (some of the other) products in the test (Kim et al. 2011, 2013). If the SE I panel needs to switch a lot between references, the 2-AFCR method has the advantage that it already performs optimally in the first few repetitions (Van Hout et. al., 2011).

When working with *d´*, it is not necessary to use similar test methods for SE I and SE II. In a SE II test with consumers that are familiar with the product, an A-Not A test should be considered as it allows direct comparison with the representation of the reference product in their memory, and it requires fewer samples than other tests. When consumers in SE II are not familiar with the reference product, training is not an option for this makes them more sensitive and less predictive of the population. Instead, a forced choice task like 2-AFCR can be used, which avoids *d´* values being underestimated because the reference is not recognised as A.

2.5 Discussion and conclusions

The reason for companies to invest in sensory and consumer research is to obtain information to base decisions on, for example, to decide whether or not to launch a new product or change an existing product. Without good sensory measures, this is merely a matter of trial and error that involves risks that can be costly. Sensory research, as laid out in this review, can help minimise these risks by selecting the most effective sensory methods. When sensory and consumer test results are expressed in *d´*, better informed decisions can be made. Using *d´* makes it possible to integrate data from different tests and gain knowledge on products differences and consumers. With relatively small investments, effects of changes in product formulations on sensory properties can be quantified with a trained sensory panel and will provide in-depth understanding of the product group. Next to this, consumers' perception of product differences can be measured for specific user groups and different countries. By relating the differences measured in the sensory panel with the sensitivities of (different groups of) consumers, the sensory panel can be developed into a powerful tool for predicting whether consumers would notice the differences. Action standards in the sensory panel can be based on the prediction of what are *meaningful differences* for target consumers. This approach allows companies to react faster to new challenges in the market, produce at higher quality, reduce the total costs of product development and minimize the failure of new product launches.

3. CAN THE SAME-DIFFERENT TEST USE A BETA CRITERION AS WELL AS A TAU CRITERION?

H-S. Lee, D. van Hout, M. Hautus and M. O'Mahony

Published in: *Food Quality and Preference, 18*, 605-613(2007)

Abstract

Using low concentration NaCl and water stimuli, judges performed same-different tests and single stimulus discrimination tests. The data were subjected to a signal detection analysis. For single stimulus judgments, a β-criterion, dividing salt vs water is assumed for calculating d′. For the same-different method, a τ-criterion, a sensory yardstick designating the degree of difference required for a 'different' judgment, is assumed. ROC analysis indicated that for single stimulus judgments, a cognitive strategy involving a β-criterion was confirmed. Yet, ROC analysis for the same-different test indicated that prior or current use of a β-criterion carried over into the same-different test for some judges, giving a mixture of τ and β-criteria.

3.1 Introduction

Sensory difference tests are used for determining whether judges can discriminate between two foods which are so similar that they can be described as confusable. Such tests are used for quality assurance, ingredient specification, product development, and studies of the effects of processing change, packaging change and storage. Sometimes they are used analytically with trained panels and such tests then come under the general heading of what has been called Sensory Evaluation I (O'Mahony, 1995a). The complimentary test is whether consumers can discriminate between foods under normal conditions of consumption, because differences that have been detected by a trained sensory panel in Sensory Evaluation I may not be detected by consumers under normal conditions of consumption. Such testing with untrained consumers then comes under the general heading of what has been called Sensory Evaluation II.

For the latter, it is generally desirable not to bias consumers by drawing attention to a particular attribute, so triangle and duo-trio tests are generally suitable. These methods lack power (Ennis, 1990, 1993) so suitable replacements would be desirable. One candidate is the same-different test, which can be more powerful than the duo-trio or triangle methods if used in a particular way (Ennis, 2004).

Each trial of the same-different discrimination test involves two samples, called the reference and comparison samples. The task requires a judge first to taste the reference sample and then taste the comparison sample, which may or may not be the same as the reference. The judge must then report whether the 'comparison' is the same as or different from the reference. This judgment has inherent response bias, so merely computing the proportion of correct responses does not give a representative measure of discrimination. However, a suitable signal detection/Thurstonian analysis can circumvent response bias and provide R-Index values (Cubero, Avancini de Almeida, and O'Mahony, 1995) or more fundamental d′ values (O'Mahony and Rousseau, 2002).

The same-different method, unlike the duo-trio and triangle methods, does not have a standard form. For the two stimuli (W and S), there are four possible orders of presentation (WW, SS, WS, SW). For the short version of the test, a test is regarded as the presentation of just one of the four possible pairs. For the long version, a test consists of the presentation of two pairs, one pair the same (WW or SS) and one pair different (WS or SW). With this test, the judge is unaware that one pair is the same and the other different. As far as the judge is concerned, he responds to each pair as if he were performing two short version tests. It is this long version of the test that modeling has indicated has more power than the duo-trio or triangle methods (Ennis, 2004). Rousseau, Meyer and O'Mahony (1998), using yoghurt stimuli, confirmed that the long version (but not the short version) of the same-different test was more powerful than the triangle method. The difference in power was only slight owing to relatively large d′ values. More published confirmations would be desirable.

Some authors have used the short version (Lau, O'Mahony, and Rousseau, 2004; Rousseau and O'Mahony, 2001; Stillman and Irwin, 1995) while others used the long version (Rousseau and O'Mahony, 2000; Rousseau, Stroh, and O'Mahony, 2002) or both (Rousseau et al., 1998; Rousseau, Rogeaux, and O'Mahony, 1999). Authors have also used a modified method in which the reference is always the same stimulus (Avancini de Almeida, Cubero, and O'Mahony, 1999; Cubero et al., 1995; Delwiche and O'Mahony, 1996).

In the present study, using short-version same-different tests d′ values were computed. A d′ value is a measure of 'effect size' (Clark-Carter, 2003). It is a fundamental measure of the perceptual difference between two stimuli, measured in units of the perceptual variation of a single stimulus. An engineer might call it a signal-to-noise ratio. Its computation involves assumptions. Yet, if the assumptions are correct, a d′ value should be independent of the method used to measure it. In this way, it is a fundamental measure. This is important for sensory evaluation because difference test measures have not been comparable between tests. The proportion of tests performed correctly with the duo-trio method cannot be compared to the proportion performed correctly for the triangle method, because their chance probabilities are different. Yet, their d' values can be compared.

What are the assumptions required of computing d′ values? The first assumption is a convenience; it is that the sensory distributions of the two stimuli (W and S) are both normal and have equal variance. The equal variance assumption would seem logical when applied to confusable stimuli; if they are so similar that they can be confused, it would be no surprise that their variances would be the same. This assumption can easily be checked and experiments appear to support it (e.g. Hautus and Irwin, 1995; O'Mahony, 1972c). A more demanding assumption made to enable the estimation of d′ is that of the nature of the cognitive strategy adopted by the judge. This always needs confirmation. For example, the computation of d′ from the triangle test assumes the adoption of a 'comparison of distances' cognitive strategy, while that for the 3-AFC assumes a 'skimming' strategy (Ennis, 1993; O'Mahony, 1995b; O'Mahony, Masuoka, and Ishii, 1994). These assumptions needed confirming. Should judges perform both 3-AFC and triangle tests, then according to the assumptions, they should perform a greater proportion of 3-AFCs correctly, but the computation of the d′ values, taking into account the appropriate cognitive strategies, should give the same results. This was confirmed by Tedja, Nonaka, Ennis, and O'Mahony (1994). A more sophisticated approach is to fit the various models to the data. For example, Irwin, Hautus and co-workers (Irwin, Hautus, and Stillman, 1992;

Irwin, Stillman, Hautus, & Huddleston, 1993; Hautus & Irwin, 1995) fitted models that assumed different cognitive strategies to ROC curves, that they obtained using different experimental procedures, and determined which ones fitted the best.

In summary, a d′ value for the same judge and the same pair of stimuli should be the same, regardless of the measurement method used, as long as the computation makes the correct assumptions. The most important assumption is the cognitive strategy. The assumptions allow the computation to circumvent the problems created by the differences in the experimental methods. It is because of this that investigation into the cognitive strategies associated with various test methods is important for sensory evaluation.

Brown (1974), when he developed his R-Index, wished for an index that was free of assumptions. It is equivalent to the index P(A), the proportion of area underneath an ROC curve formed by connecting the points with straight lines (Green and Swets, 1966). It has been reviewed by O’Mahony (1992). The computation of P(A) (R-Index) does not use assumptions to circumvent the differences between the experimental methods used to measure it. Accordingly, it is prone to vary with experimental method. For example, P(A) for the same-different test will vary with the cognitive strategy used by the judge (see below). Also, it was shown that an R-Index obtained by ranking is higher than one obtained from rating data, because of the forced-choice nature of ranking (Ishii, Vié, and O’Mahony, 1992; O’Mahony, Garske and Klapman, 1980. Thus, the R-Index or P(A) can be seen as a measure of ‘performance’ rather than a fundamental measure of difference.

To explore the cognitive mechanisms associated with the same-different test, it is important to consider response biases and the criteria associated with difference tests. Consider a judge being given a set of confusable stimuli (W and S) and being required to report their identities (say “W” or “S”). This procedure has been called the yes/no task (Green and Swets, 1966). As the stimuli are confusable, the decision as to whether a stimulus is ‘W’ or ‘S’ can be difficult to make. His response will be the result of how well his receptors can distinguish between the two sensory signals elicited by ‘W’ and ‘S’ and also where he ‘draws the line’ between the sensations he would categorize as coming from ‘W’ and those he would categorize as coming from ‘S’ (Green and Swets, 1966; O’Mahony, 1992, 1995b).

Depending on where he ‘draws his line’, he may be biased and more willing to categorize his sensations as ‘W’ or biased towards categorizing his sensations as ‘S’: hence the term ‘response bias’. The ‘line’ has a technical name. In food science, it is called the β-criterion (Rousseau, 2001; Rousseau et al., 1998). The cognitive strategy that uses this criterion has been called the β-strategy (Rousseau, 2001). Data from such an experiment can be used to produce ROC curves and to estimate d′ values (Green and Swets, 1966). If a β-strategy were to be used in a same-different test, it would require the judge to identify each stimulus in the pair independently and then decide whether they fell on the same side or on different sides of the β-criterion line. In psychology, this strategy has been called the 'independent observation model' (Macmillan and Creelman, 1991) and the 'optimal' strategy or decision rule (Noreen, 1981; Irwin and Francis, 1995). The 'optimality' of the β-strategy relates to the fact that P(A) (R-Index), the proportion of area below the ROC curve for a given value of d′, turns out to be larger for this strategy than for the τ-strategy.

To digress briefly, it can be seen that psychology and sensory food science use a different set of technical terms and symbols. Because it is sometimes necessary for the food scientist to explore the psychological literature, it is as well to be aware of these differences. For example, the psychologists Green and Swets (1966) denote the β-criterion by the symbol 'k'. To add to the confusion, they use the symbol 'β' to denote something completely different: the likelihood ratio at the criterion point. One crosses interdisciplinary boundaries with care.

In a same-different test, a judge has two possible cognitive strategies at his disposal (Hautus and Irwin, 1995). Besides the β-strategy, the judges can use a second strategy. This involves the use of a τ-criterion (Rousseau, 2001; Rousseau et al, 1998). A τ-criterion is concerned with how different two stimuli need to be, to be reported as 'different'. It can be visualized as a sensory yardstick. If sensations elicited by the two stimuli in the same-different test are more different than the yardstick, the stimuli will be reported as different; if not, they will be reported as the same (Irwin and Francis, 1995; Irwin et al., 1993; Rousseau, 2001; Rousseau et al., 1998). In psychology, the τ-criterion has been called a k-criterion (Macmillan and Creelman, 1991; Macmillan, Kaplan, and Creelman, 1977). The cognitive strategy that utilizes the τ-criterion will here be called the τ-strategy. In psychology, it has been called the 'differencing model' (Macmillan and Creelman, 1991) or the 'sensory difference decision rule' (Noreen, 1981).

For a standard yes/no task, in which a judge is presented with either S or W and must indicate which sample was presented, the β-strategy is the only available strategy that can be adopted. For this case, d′ can be calculated from the standard yes/no ROC curve (Green and Swets, 1966). If the sensation distributions are normal with equal variance, the ROC curve will be symmetrical about the negative diagonal and a z-plot ROC will be a straight line with a slope of unity. The regular ROC curve will lose its symmetry and the slope of the z-plot will deviate from unity if the variances of the two distributions are not equal; the slope will be equal to the ratio of the standard deviations of the two distributions.

This standard ROC analysis used for the yes/no task cannot be applied to the same-different task. A different approach is required to compute d′. There is one approach if a β-strategy is used and another if the τ-strategy is adopted. If a β-strategy is used for the same-different test, the ROC will be symmetrical, as is the ROC for the standard yes/no method with equal variance. However, the same-different ROC will not be exactly the same shape as that for the standard yes/no model. It will be symmetrical, yet it will have a higher proportion of area, P(A), under the curve. If the two curves were to be overlaid, the same-different ROC would be seen to bulge out further than the yes/no curve. Yet, the two curves would meet where they intersect the negative diagonal and, of course, would also meet at the ends of the curve. So the same-different curve can be described as rising up more quickly and then flattening out to meet the lower yes/no curve at the negative diagonal. Yet, even though the P(A) values would be different, the d′ values, taking into account the different cognitive strategies, should be the same.

If a τ-strategy is used, the ROC will be asymmetrical as is the ROC for the standard yes/no model with unequal variance. Again, the two ROCs will not be the same shape, however the differences are more complex than those for the β-strategy ROC given above. Again, for the same-different model, the proportion of area under the curve will be greater than for the yes/no task with unequal variance, yielding higher P(A) (R-Index) values.

To summarize, ROC curves for same-different method using the β-strategy and for the yes/no method (equal variances) using the β-strategy are both symmetrical. Yet P(A) for the same-different method using the β-strategy is greater. The ROC curves for the same-different method using the τ-strategy and for the yes/no method using the β-strategy (unequal variances) are both asymmetrical but P(A) is greater for the same-different method. However, for all these procedures the computed d′ values should be the same, if the computation takes account of the appropriate cognitive strategy. In addition, Hautus and Irwin (1995) indicated that P(A) values were greater for the yes/no method (β-criterion) than for the same different method (τ-criterion); again computed d′ values will be the same if the appropriate cognitive strategy is assumed..

Because it is important to know the cognitive strategy being used in a difference test to be able to compute d′, it is important to investigate such strategies. One approach to determining the cognitive strategy is simply to interview the judge (Tedja et al. 1994) or to require the judge to ‘think aloud’ (Wong, 1997). Because judges may not always be aware of their cognitive strategy, a second approach is to examine the ROC curve obtained for the judge. The present study relied on this latter approach although the occasional judge would ‘think aloud’ and was not discouraged.

The fact that the ROCs for the same-different β-strategy and the standard yes/no task with equal variances are both symmetrical, and those for the same-different τ-strategy and the standard yes/no task with unequal variances are both asymmetrical, leads to a simplified ROC analysis to determine whether a τ-criterion is being used for the same-different task. Simply fit the standard yes/no ROC to the same-different data to determine whether or not the ROC curve is symmetrical. If the variances of the two sensory distributions can be assumed to be the same, the standard yes/no ROC curve will be asymmetric for a τ-strategy (Irwin et al., 1993; Hautus and Irwin, 1995). Also, the slope of the z-plot ROC will be greater than unity. This shortcut method is only useful for determining the cognitive strategy being used. It will not provide a legitimate estimate of d′. The computation of the d′ value is more complex. It requires fitting the appropriate same-different model of the ROC to the data. Alternatively, for the τ-strategy, d′ can be estimated without undertaking an ROC analysis, by using Ennis’s method of computation given in O’Mahony and Rousseau (2002).

In psychology, there has been considerable discussion regarding cognitive strategies or decision rules involved in the same-different method (for example, Dai, Versfeld, and Green, 1996; Irwin and Hautus, 1997; Noreen, 1981, Sorkin, 1962). Kaplan, Macmillan, and Creelman (1978) provided tables of d′ for the same-different test. Irwin and Francis (1995) noted that different cognitive strategies were adopted for visual stimuli, depending on the complexity of the stimuli. A set of judges used a β-strategy for complex stimuli (kanji: Japanese system of writing using Chinese characters) while with what appeared to be a separate group of judges, a τ-strategy was adopted by two out of three judges for more simple stimuli (colors). They also studied same-different tests, where the stimuli were judged as conceptually the same or different (natural vs manufactured items), rather than physically so (Irwin and Francis, 1995; Francis and Irwin, 1995); their results supported a β-strategy.

Yet, it is in the work of Irwin, Hautus, and their co-workers, who considered taste and food stimuli, that the considerations of cognitive strategy become more relevant to sensory evaluation. Irwin et al. (1992) reviewed ROC curves and some of the drawbacks of R-Index measurement for methods that induced a β-criterion. Irwin et al. (1993) demonstrated how ROC curves, derived from same-different tests for orange drinks, were best fitted assuming a cognitive strategy that used a τ-criterion. The same result was obtained by Stillman and Irwin (1995) using a raspberry flavored drink. These data were supported by same-different experiments with auditory stimuli (Hautus, Irwin, and Sutherland, 1994). From these studies, it would seem that in the same-different test, judges tend spontaneously to adopt a τ-cognitive strategy. Irwin, Hautus and co-workers went on to examine bias and the interpretation of areas under ROC curves for the same-different test (Irwin, Hautus, and Butcher, 1999; Irwin, Hautus, and Francis, 2001). However, the most relevant paper to the present study is that of Hautus and Irwin (1995).

Hautus and Irwin used the signal detection rating procedure to determine how well judges could distinguish between milks of different fat content. In a yes/no task, a random order of milk stimuli was tasted and judges had to report which milk they tasted and rate the sureness of their responses on a 6-point category scale. Values of d′ were calculated and symmetrical ROC curves obtained (using the standard yes/no model), indicating that the two sensory distributions had equal variance. In a second experiment, the same stimuli were discriminated using a same-different test, again with sureness ratings on a 6-point scale. The ROC curves obtained were asymmetric and the d′ values calculated assuming a τ-criterion, agreed with those in the first experiment. The d′ values were fairly close to threshold, so it was not possible to tell whether an analysis assuming a beta criterion would have fitted the data better. However, from past research, it would seem unlikely.

In Experiment I for the present study, the goal was to determine whether it was possible for judges to use a β-strategy for a same-different taste test. Accordingly, judges performed short-version same-different tests under two protocols. For one protocol, the experimental conditions were set up to favor a τ-strategy while for a second protocol they favored a β-strategy. ROC curves were examined to determine which strategy was actually used for each protocol. Should the β- and τ-strategies be used in their appropriate protocols, then it may be expected that computed d′ values computed from both protocols and also from some additional 2-AFC tests should correspond.

3.2 Experiment I

3.2. 1. Materials and methods

Judges

Eleven judges (3 M, 8 F; age range 21-62 yrs.), students, staff and friends at UC Davis, participated in the experiment. Judges were required to fast, except for water, for at least 1h prior to testing. Five had participated in taste psychophysical experiments beforehand, six had not.

Stimuli

Stimuli consisted of low concentration NaCl solutions (0.5-5.0 mM, depending on the judge's sensitivity) to be discriminated from purified water. The NaCl solutions (S) were prepared by dissolving reagent grade NaCl (Mallinckrot Inc., Paris, KY) in Milli-Q purified water. The Milli-Q purified water was deionized water fed into a Milli-Q system involving ion exchange and activated charcoal (Millipore Corp., Bedford, MA). The resulting purified water had a specific conductivity of $< 10^{-6}$ mho/cm and a surface tension ≥ 71 dynes/cm. The purified water was used as the water stimulus (W).

Stimuli were dispensed in 10 ml aliquots using both Repipet Adjustable Dispensers (Labindustries Inc., Berkeley, CA) and Oxford Adjustable Dispensers (Lancer, St Louis, MO) in plastic cups (1oz portion cups, Solo Cup Co., Urbana, IL). All stimuli were served at constant room temperature (21-24 °C), on white plastic cutting trays. Stimulus concentrations ranged 0.5-5.0 mM (0.5-3.0 mM, 2 judges; 1.0-3.0 mM, 7 judges; 3.0-5.0 mM, 2 judges).

Procedure

Each judge performed 96 same-different tests under each of two protocols. In the beta-protocol, conditions were arranged to encourage the use of a β-criterion. In the tau-protocol, a τ-criterion was encouraged. The tests were performed over 4 separate sessions on separate days (24 tests under each protocol per session, total=96 per protocol).

For what will be called the beta-protocol, judges first performed a warm-up procedure (Dacremont, Sauvageot and Duyen, 2000; O'Mahony, Thieme and Goldstein, 1988; Pfaffmann, 1954; Thieme and O'Mahony, 1990). The warm-up consisted of tasting alternately water and salt stimuli, so the judge could discover the sensory signals that denoted each one. In other words, the judge was establishing a β-criterion, differentiating between water and salt. Each judge tasted at least 5 of each stimulus and more if desired. After the warm-up, judges then performed 24 same-different tests which had been modified to promote the use of a β-criterion. The judge, when presented with the pair of stimuli, was required to report whether each stimulus was water or salt. If both were reported as water or as salt, the test response was scored as "same". If one of the stimuli was reported as salt and the other as water, the test response was scored as "different". Judges were also required to say whether they were sure or unsure of their pair of judgments.

For what will be called the tau-protocol, a modified warm-up procedure was used. Judges were presented with two pairs of stimuli. The first pair consisted of two water stimuli and the second pair consisted of water followed by salt (W-W, W-S). Judges tasted at least five of each of these pairs and more if desired. Next, two further pairs (S-S, S-W) were tasted in the same way. The goal of this warm-up was for judges to discover the signals indicating same and different stimulus pairs and thereby to establish a τ-criterion, indicating the degree of difference required for a "different" response. After this modified warm-up, judges then performed 24 regular same-different tests. Again, judgements of "sure"/"not sure" were added. The judges were discouraged from identifying the stimuli and instructed only to pay attention to whether they felt the stimuli were the same or different. Subjective responses indicated that judges could perform according to each protocol; four judges who reported difficulty with these tasks were eliminated.

After establishing rapport, and taking demographic details, the experimenter instructed the judge to take at least 6 purified water mouthrinses to clean the mouth. Judges then performed the warm-up for the specific protocol and the same-different tests without any interstimulus rinsing. After a further 6 mouthrinses, the warm-up procedure and same-different tests were performed for the other protocol.

Also included in each session were twelve 2-AFC tests. Before these tests, judges took 6 mouthrinses, and performed a warm-up with at least 5 pairs of water and salt stimuli. No further mouthrinses were taken after the initial six. The 2-AFC tests were performed either at the beginning or the end of an experimental session. There were thus, four possible orders of presentation for the beta-protocol, tau-protocol and 2-AFC protocol within an experimental session. These were: beta/tau/2-AFC, 2-AFC/tau/beta, tau/beta/2-AFC, and 2-AFC/beta/tau. The four orders of presentation were all used for each judge over the 4 experimental sessions. The order of presentation of stimuli within a given test, was chosen randomly for one judge and the reverse order was employed for the next judge. For the third judge, a separate random order was chosen, and so on. Subjects responded verbally. Experimental session lengths ranged 15-43 min.

Prior to the first experimental session, judges had a training session. Judge sensitivity was determined using 2-AFC tests. Judges who could not discriminate between purified water and 5mM NaCl were eliminated. Judges practiced using the beta and tau-protocols and those who experienced difficulty using the two cognitive tasks were eliminated based on their subjective reports. From the results of this practice session, it was decided that 5mM NaCl should be used for the same-different tests in the main experiment. However, it was necessary to use a lower concentration (3mM) for pairs where NaCl was tasted after purified water. This was because of sensitization to NaCl caused by a lowering in the adaptation level by the water stimulus (Bartoshuk, 1968, 1974, 1978; Halpern, 1986; McBurney and Pfaffmann, 1963; O'Mahony, 1972a, b, 1979; O'Mahony and Godman, 1974). Thus, the session used 5mM NaCl, 3mM NaCl and water. If a ceiling effect was encountered or a more sensitive judge was tested, the concentrations were reduced to 3mM and 1mM NaCl. A further reduction to 1mM and 0.5mM NaCl was found necessary for two judges. Nine judges started with 3 and 1mM, while two started at 5 and 3 mM. Because the goal of the experiment was only to compare performance on the same-different test under the beta and tau protocols, the variations in concentration for these judges did not invalidate the experiment because exactly the same concentrations were used an equal number of times under each protocol.

3.2.2 Results and discussion

Mean d′ values, computed from the 11 judges tested in Experiment I, are shown in Table 3.1 using a variety of models.

Table 3.1 Mean d’ values computed using various analyses from the results of Experiment I (N=11)

	Same-Different test	
	Tau-protocol	**Beta-protocol**
From 2-AFC method	From same-different judgments using computation that assumes τ-criteria	From salt vs water judgments using computation that assumes β-criteria
1.86 [a]	1.82 [a]	1.54 [a]

[a] means were not significantly different (p>0.05).

The values were computed using the IFPrograms software (Institute for Perception, Richmond, Virginia). Significant differences between these means were computed using ANOVA and LSD tests (p<0.05). The first (1.86) was computed from the 2-AFC tests and can also be derived from Tables (Ennis, 1993). The second (1.82) value was derived from the same-different test performed in the tau-protocol, using a computation (degree of difference program, IFPrograms) that assumed a τ-criterion (O’Mahony and Rousseau, 2002). The third mean (1.54) was derived from salt vs water judgments from the beta-protocol using a computation (scale program, IFPrograms) that assumed a β-criterion (Kim, Ennis and O’Mahony).

To investigate whether τ or β-strategies were used in the same-different tests, ROC curves were fitted to the data in the different/same matrices using maximum-likelihood estimation (Hautus et al., 1994). The data were pooled across judges. This has drawbacks for estimating sensitivity (Macmillan and Creelman, 2005). However, if sensitivity is not the main area of interest, then pooling data can give a stricter test of the models under investigation since the variability of each point on the ROC curve can be dramatically reduced. The pooled data for the tau and beta protocols are given in Table 3.2.

Table 3.2 Results of ROC analyses for pooled data for tau and beta-protocols in Experiment I

Protocol	Fitted cognitive strategy					
	τ-strategy			β-strategy		
	d′	χ^2	p*	d′	χ^2	p
tau	1.72	**2.43**	**0.296**	1.40	7.74	0.021
beta	2.38	**1.74**	**0.419**	1.93	8.51	0.014

*p = probability that data arose from the model fitted given that the model is correct.

Small χ^2 and large p signifies a good fit.

Bold and underlined χ^2 and p values indicate best fitted strategy for ROC curve.

The table indicates the best fitting value of d′, the goodness-of-fit statistic (χ^2), and the probability that the data arose from the model fitted, given that the model was correct (p). Smaller values of χ^2, and larger values of p, indicate a relatively better fit. It can be seen from the table for both χ^2 and p, the results suggest that the best overall fit in both protocols was the τ-strategy.

Considering the three d′ values from Table 1, the fact that they were not significantly different would confirm signal detection/Thurstonian theory. Each value had been computed taking into account the appropriate cognitive strategy. Yet, the value for the beta-protocol was rather low. However, the ROC analyses indicated that a τ-strategy rather than a β-strategy was being used for this protocol. In this case, because P(A) is smaller for the τ-strategy, a computation assuming a β-strategy would underestimate d′.

There are additional possibilities. Because the sureness judgments for the beta-protocol were given for pairs of stimuli rather than for each individual stimulus some boundary variance may have been introduced. Boundary variance is variance introduced because judges vary in their β-criteria (boundaries) between what should be reported as “water” or “salt” and their judgements of “sure” and “unsure”. Such added variance would depress the value of d′.

Thus, from the pooled data, it may be concluded that this experiment failed to demonstrate that a β-strategy would be adopted by judges in a condition favorable to its adoption (beta-protocol). Yet, this conclusion is not so clear when data from individual judges are examined. For the tau-protocol, the data computed for each individual judge indicated that six judges had a better fit with the β-strategy while only five had a better fit with the τ- strategy. Surprisingly, for the beta protocol, only two had a better fit for the β-strategy while nine had a better fit for the τ-strategy. There are various explanations for this mix of strategies. It may be that the four category sureness scale used to generate the data did not give sufficiently accurate ROCs. It may be that judges carried over their strategies from one protocol to another. It might also be that some of the data could be the result of the intrusion of a third unexpected strategy. Unexpected strategies have been reported before (Tedja et al., 1994). However, it must be stressed that the results for individual judges must be interpreted with care. They are more prone to sampling error compared with the pooled results.

Notwithstanding, the fact that some judges might have used a β-strategy requires more attention. Accordingly, the possibility of the use of a β-strategy was investigated further in Experiment II

3.3 Experiment II

The goal of this experiment was to use an improved experimental design to determine whether a prior set of single stimulus judgments, requiring a β-strategy, could affect the choice of strategy for a subsequent same-different test. To enable better fitting of ROC curves than in Experiment I, judges were required to perform more tests and responses were given on a 6-point rather than a 4-point scale. Also, to check that prior judgments of single stimuli did conform to the β-strategy, independent sureness judgments were made for each stimulus to allow the construction of ROC curves. These curves which were not available from the first experiment because of the sureness judgment regime used, were available for comparison with same-different ROCs,

3.3. 1. Materials and Methods

Judges

Four judges (4 F; age range 22-44 yrs.) who had participated in Experiment I were available to return for more intensive re-testing.

Stimuli

The stimuli were the same as in Experiment I with NaCl concentrations adjusted to produce d′ values in the range 1.8–2.5. This avoided ROC curves being too close to the positive diagonal, where it is not easy to distinguish between ROC curves generated by β and τ-criteria. Stimulus concentrations ranged 1.0-5.0 mM (1.0-3.0 mM, 1 judge; 3.0-5.0 mM, 3 judges). Because the goal was to determine the shape of the ROC curves and not make estimates of d′, combining data over sessions with slightly different signal strengths did not invalidate the data.

Procedure

Each judge performed 24 tests per experimental session. In each experimental session a test consisted of tasting a pair of stimuli. The judge was required to rate the first stimulus as being ‘salt’ or ‘water’ using three levels of sureness: “sure” vs “not sure” vs “I do not know but I will guess” resulting in a 6-point category scale. This is the signal detection rating procedure described by Green and Swets (1966). The judge was then required to rate the second stimulus in the same way. Finally, she was required to judge whether the two stimuli were the same or different, using the three levels of sureness for this same-different judgment. All four possible pairs were used (WS, SW, WW, SS).

To check whether the judge’s daily variation in sensitivity was in the range that was advantageous for fitting same-different ROC curves (d′=1.80-2.50), a set of six 2-AFC tests was performed prior to and after the end of the testing. Judges began by taking at least 6 mouthrinses. They then performed a warm-up as in Experiment I for the beta-protocol. Immediately after this, they performed six 2-AFC tests. They were then offered the option of a further warm-up if desired before beginning the 24 tests. After the 24 sets of same-different tests, judges were given a further six 2-AFCs without prior warm-up just to double-check the sensitivity change after the tests. Judges performed 8 sessions, giving a total of 192 same-different tests.

If the initial six 2-AFC tests indicated that the judge did not have the required sensitivity, a further six 2-AFCs were performed as a check before any decision was made about abandoning that experimental session. If the data analysis for a given session indicated that the judge’s sensitivity was not in the specified range, the session was rescheduled and repeated. The number of abandoned sessions ranged 2-8 per judge. As the sensitivity of judges changed with practice, the NaCl stimulus concentrations were varied to keep the judges within the required sensitivity range. Experimental session lengths ranged 20-45 min. Generally, experimental sessions were performed on separate days although some judges chose to perform more than one session per day. Testing ranged over 7-11 days per judge. Other details of the procedure were as for Experiment I.

3. 3. 2. Results and discussion

ROC curves were fitted to data obtained from the first stimulus and the second stimulus and the same-different judgments. They were fitted using maximum-likelihood estimation (Hautus et al., 1994). The results are shown in Table 3.3.

Table 3.3 Results of ROC analyses for same-different test with single stimulus and same-different judgments in Experiment II.

ROC analysis for single stimulus judgments (β-strategy)								
	First stimulus				Second stimulus			
Judge	d′	Slope	χ^2	p	d′	Slope	χ^2	p
A	2.20	1.67	**4.14***	**0.246**	1.43	0.94	**1.43**	**0.698**
B	2.97	2.26	**5.55**	**0.136**	1.70	2.40	**5.64**	**0.130**
C	2.53	1.18	7.66	0.054	1.86	0.79	**3.00**	**0.392**
D	2.26	1.49	**5.35**	**0.148**	1.30	1.99	**0.21**	**0.977**
Pooled	2.30	1.69	21.8	<0.001	1.50	1.35	1.28	0.733

ROC analysis for same-different judgments						
Judge	τ-strategy			β-strategy		
	d′	χ^2	p	d′	χ^2	p
A	2.29	**3.89****	**0.421**	1.80	8.79	0.066
B	2.76	7.19	0.126	2.19	**3.15**	**0.532**
C	2.47	6.39	0.172	1.96	**5.20**	**0.267**
D	2.27	**3.12**	**0.537**	1.88	4.26	0.372
Pooled	2.38	6.62	0.157	1.90	11.16	0.025

*statistic and probability for goodness of fit (β-strategy) for the single stimulus judgment in bold

** statistic and probability for best fitted strategy for same-different judgments for ROC curves in bold and underlined.

Considering the results for the single stimulus judgments, the standard β-strategy (yes/no) model was fitted. The fit of the model for the ROC curves was good in all cases except for the judge C with the first stimulus. This indicated that the judges were, at this point, decision making in terms of β-criteria. The slopes for both stimuli tended to be greater than unity, indicating a larger variance for the 'noise'; values less than unity were not significantly so. For all judges, d′ for the first stimulus was higher than for the second stimulus. An examination of data indicated that this was mainly due to errors when both stimuli were salt. This is expected from earlier research on sequence effects and is predicted by the Conditional Stimulus and Sequential Sensitivity Analysis models of discrimination (Ennis and O'Mahony, 1995; O'Mahony and Goldstein, 1987; O'Mahony and Odbert, 1985; Tedja et al., 1994; Vié and O'Mahony, 1989). Thus, the first stimulus is the better representative of the use of a β-strategy for single stimulus judgments.

Considering the fit for the same-different judgments at the bottom of the table, the χ^2 and p values were consistent with the judges A and D using the τ-strategy while judges B and C used the β-strategy. It may be hypothesized that judges B and C carried over their β-decision making strategies into the same-different test while judges A and D did not. This would confirm the results from Experiment I where some judges used a τ-strategy while others used a β-strategy. This is interesting because, as outlined in the Introduction section, research generally indicates a τ-strategy for same-different judgments. The present research indicates that the choice of τ-strategy can be interfered with by prior β-strategy decision making. Further research should broaden the knowledge on strategy choice for the same-difference test.

Three judges (A, B, and C) were consistent in their choice of strategy for same-different judgments in Experiments I and II. The fourth inconsistent judge (D) did not fit either strategy well in Experiment I. Values of d′ for the first single stimulus and the same-different test favoring the τ-strategy (judges A and D) showed good agreement. Yet, for judges B and C (β-strategy), agreement was not good, although sometimes complicated by low p-values.

In conclusion, although the same-different method with simple sensory stimuli has generally been shown to use a τ-criterion, the present research indicates that if judges were involved in prior or simultaneous decision making using a β-criterion, a β-decision rule or cognitive strategy may be adopted by some judges for the same-different test. This adds to the work of Irwin and Francis (1995) and Francis and Irwin (1995) who found that for conceptual same-different judgments and judgments of complex stimuli, a β-strategy best fitted ROC data.

One may speculate that judges, if exposed to the same or similar stimuli over a period of time, may begin to categorize stimuli and begin to use a β-criterion for making their same-different judgements. Also, it may be hypothesized that what applies to the same-different test in terms of cognitive strategies may also apply to the versions of the

A-Not A method (ASTM, 1968; Peryam, 1958; Pfaffmann, 1954). These are all topics for further research.

4. SENSORY DIFFERENCE TESTS FOR MARGARINE: A COMPARISON OF R-INDICES DERIVED FROM RANKING AND A-NOT A METHODS CONSIDERING RESPONSE BIAS AND COGNITIVE STRATEGIES

H-S. Lee, D. van Hout, and M. O'Mahony

Published in: *Food Quality and Preference, 18*, 675-680 *(2006)*

Abstract

Sensory difference tests were performed between 6 margarine products, a standard vs 5 other products. Three testing protocols were used. The first protocol was simple ranking. The second protocol was the A-Not A method where a single standard was presented beforehand and which could be retasted during testing. The third protocol was the A-Not A method where all products were presented beforehand but could not be retasted during testing. R-Index values were computed for each protocol. Ranking gave the highest R-Index values while the A-Not A method, where only a single standard was presented prior to testing, gave the lowest R-Index values. R-Indices were calculated by averaging indices from individual judges and also by pooling data from all judges. Differences between these computations only occurred for the A-Not A method where all the products were presented prior to testing. Differences were explained in terms of the forced-choice nature of ranking, boundary variance, concept formation and differences in cognitive strategies involving tau and beta-criteria.

4.1 Introduction

Sensory difference tests are used for discriminating between two confusable food stimuli or other products with sensory attributes. Such tests are used for reformulation, quality control, product development, ingredient specification, shelf-life, cost reduction, packaging studies etc. One such test is the A-Not A method, sometimes called the single stimulus method. Although A-Not A test is not the most common test, it is used by food industry. This method was first introduced to food science by Pfaffmann, Schlosberg and Cornsweet (1954). For this protocol, a product (A) is presented to the judge several times at the start of an experimental session so that the judge can become familiar with it. Then, a series of two products, 'A' and a slightly different product to be discriminated from 'A' (Not A) are presented in random order. The judge has to respond by stating which products are 'A' and which are 'Not A'. During the test session, the judge is given 'A' at various intervals, knowing its identity, as a reminder.

Peryam (1958) later described the test in the same way except that he stated that not only 'A' but sometimes 'Not A' could also be presented at the beginning of the test for familiarization. Although both products were removed before testing, 'A' could be presented to the judge during the experimental session as a reminder, as described by Pfaffmann.

Meilgaard, Civille and Carr (1991) also described the test with both products being presented beforehand for familiarization. However, in their version of test, no reminders are given. They did mention that although usually only one 'Not A' product is presented during testing, sometimes more than one 'Not A' product could be presented. In such a case, all possible 'Not A' products should be presented prior to the test. ASTM (1968) described the test indicating that the two products (A and Not A) are given beforehand; they are vague on further details. Lawless and Heymann (1996) point out that there have been various versions of the A-Not A method. It would appear that there is no agreed standard A-Not A method; as long as the general procedure is followed, the method is given its name. Because of this, it is always better to describe in detail the methods being used. The different methods have the potential to change the cognitive strategy being used. If there were changes in cognitive strategy, comparisons of the discrimination indices between methods would be problematical. Because of this it is worth gaining understanding of these effects.

The problem with the A-Not A method is that like the same-different method, it has inherent response bias (O'Mahony and Rousseau, 2002). Specifically, for the A-Not A method, a judge's response will not only depend on whether his sensory system is sufficiently sensitive to discriminate between the products 'A' and 'Not A', but also depend on his willingness to report a different product as 'Not A'.

Because of such response bias, merely counting the proportion of correct responses in a A-Not A method is a biased measure of perceived difference. However, response bias can be circumvented by using a signal detection/Thurstonian analysis to calculate an index of difference such as d′ (Macmillan and Creelman 2005; O'Mahony and Rousseau 2002). An R-Index computation could also be used (Delwiche and O'Mahony 1996). For such computations, judges would generally add sureness judgments to their 'A' versus 'Not A' judgments (Brown, 1974).

For the computation of d′, it is necessary to know the cognitive strategy being used in the A-Not A test (O'Mahony and Rousseau, 2002; Lee and O'Mahony, 2004; O'Mahony, Masuoka and Ishii, 1994). There are two logical possibilities. Firstly, the A-Not A method could be treated as a version of the Yes-No procedure in signal detection which implies the use of a beta-criterion (Green and Swets, 1966). For this it is assumed that there are only two products in the test. In this case, the judge would hold two categories in his head, one corresponding to 'A' and the other to 'Not A'. The boundary between the two categories would be the beta-criterion. The judge would assign the test products to either category and respond accordingly. Recent neuropsychological research has reported that the A-Not A method using multiple stimuli (more than one type of 'Not A') also utilizes a beta-criterion (categorization) (Casale, Ashby and Standring, 2005).

Secondly, the A-Not A method could be considered as an extension of the same-different method. This method is assumed to use a tau-criterion., although there are a few exceptions (Lee, van Hout, Hautus and O'Mahony 2006; Irwin and Francis 1995; Francis and Irwin 1995), The tau-criterion is a degree to which two stimuli must differ, to be reported to be different. If the two products are more different than the tau-criterion, they are reported as different. However, if they are not more different than the tau-criterion, they are reported as the same. Accordingly, if the product to be tested is perceived as more different from 'A' than the tau-criterion, the judge will report it as 'Not-A'. If it is not perceived as more different, it will be reported as 'A'.

Should the A-Not A method be used in a situation where there are several 'Not A' products, an alternative procedure might be to use a simple ranking method. Judges would rank the products in terms of their similarity to 'A'. Designating 'A' as the 'noise', the degree of difference between 'A' and the various test products (signals) could be computed using an R-Index analysis (Brown 1974; O'Mahony 1992). The A-Not A method is essentially a rating or categorization procedure from which R-Index values can be computed to represent the degree of difference between 'A' and the various 'Not A' products. The ranking procedure would also give R-Index values representing the difference between 'A' and the various 'Not A' products. However, because of its forced choice nature, ranking tends to give higher R-Index values than those calculated from rating or categorization (O'Mahony, Garske and Klapmann, 1980; Ishii, Vié and O'Mahony, 1992).

Prior research indicating that ranking gives higher R-Index values than a simple rating or categorization procedure was performed using a simple model system (O'Mahony et al., 1980; Ishii et al., 1992). The goal of the present experiment was to investigate whether ranking gave higher R-Index values than a more complex rating or categorization procedure, namely the A-Not A method. Specifically, the goal of this experiment was to compare a ranking protocol with two A-Not A protocols. In all cases, there was more than one 'Not A' product. For one protocol, only 'A' was tasted beforehand for familiarization. For a second protocol, 'A' and all 'Not-A' products were given beforehand for familiarization. For a third protocol, the products were simply ranked.

4.2 Materials and methods

4.2.1. Judges

Seven experienced female panelists (age range, 45-61 yrs.) were tested. Their experience of participating on sensory panels for testing margarines ranged 5-12 years. All were familiar with the A-Not A and ranking methods.

4.2.2. Stimuli

Six commercial margarines were obtained from the local supermarkets in Vlaardingen, Holland. These were: (A) Halvarine (Gouda's Glorie, Zeewold, NL), (B) Havarine (Perfekt, Beesd, NL), (C) Bona, (Unilever Netherlands, Rotterdam, NL), (D) Volle Pond (Gouda's Glorie, Zeewold, NL), (E) Harvarine (C.I.V. Superunte B.A., Beesd, NL) (F) Sunflower (Gouda's Glorie, Zeewold, NL). For the purposes of this article, these products will be referred to by their corresponding letters 'A' to 'F'. All products were presented in 50 ml white plastic lidded cups under red light to minimize any color and reflectance differences. To sample the product, judges removed the lid and used separate plastic teaspoons for each tasting. Products were tasted and swallowed. Products were served chilled (5 °C) having been stored in a fridge until 5 min before serving. Between tastings, judges rinsed ad-lib with room temperature de-mineralized water (23 - 24 °C). Before beginning each of the three protocols, judges were allowed to eat Barber crackers (the horizon Biscuit Company Ltd., England) if desired; after this all judges then rinsed at least five times.

4.2.3. Procedure

Judges performed difference tests between the margarines using three separate protocols.

For the first protocol: 'ranking', product 'A' was presented to the judges as the standard. Judges were able to taste the standard as much as desired until they felt they had become familiar with its sensory characteristics (at least 4 teaspoonfuls). They were then given products 'A' to 'F' simultaneously and instructed to rank them in order of similarity to the standard. During testing, the standard and products 'A' to 'F' could be retasted as often as desired.

For the second protocol, a version of the A-Not A method was used. As before, product 'A' was presented to the judges as the standard. Again, judges were able to taste the standard as much as desired until they felt they had become familiar with its sensory characteristics (at least 4 teaspoonfuls). They were then given products 'A' to 'F' individually in random order, counterbalanced over sessions and required to report whether the products tasted the same or different from the standard. Responses were given in terms of six categories as follows: "same sure", "same not sure", "don't know, but guess it's the same", "don't know, but guess it's different", "different not sure", "different sure". During testing, the standard 'A' could be sampled as much as desired. For the purposes of this article, this protocol will be referred to as 'A-Not A: single'.

For the third protocol, a different version of the A-Not A method was used. As before, product 'A' was presented to the judges as the standard. However, this time products 'B' to 'F' were also presented simultaneously with 'A'. Judges were able to taste all these products as much as desired until they felt they had become familiar with the sensory differences between the standard 'A' and the products 'B' to 'F'. They were then given products 'A' to 'F' individually in random order and required to report whether they tasted the same or different from the standard. During testing, judges were not allowed to retaste the standard 'A' at will. For the purposes of this article, this protocol will be referred to as 'A-Not A: multiple'.

Judges performed all three protocols in a single session. They performed two sessions per day, lasting approximately 2 ½ hours, for a total of seven days (total 14 sessions). The order of presentation of the protocols was counterbalanced over sessions. There was a week interval between the first two days of testing. After a period of 10 months, testing the final 5 days was resumed at one week intervals. This schedule was determined by the limited availability of the trained taste panel for experimental work. However, examination of the data indicated that this unusual schedule did not adversely affect judges' performance.

4.2.4. Statistical analysis

R-Indices were computed in two ways: Firstly, for each product ('B' to 'F' as signals and 'A' as noise) and each protocol, R-Indices were computed individually for each judge (number of signals/noise = 14 per judge). Mean R-Indices, across judges but within protocols, were then calculated. For the second analysis, for each product ('B' to 'F' as signals and 'A' as noise) and each protocol, data for all judges were pooled onto a single response matrix and a single R-Index was computed. (number of signals/noise = 98 = 7 judges x 14 sessions).

4.3 Results and Discussion

The computed R-Index values for products 'B' to 'F' for the three protocols are given in Table 4. As noted above, the two R-Index computations involving averaging judges' individual data and pooled data are also shown, as are means for all the products.

Table 4. R-Index values indicating differences between margarine products derived from ranking and two A-Not A methods, using two ways of combining data from individual judges.

Method of combining judges' data	Products	Protocols			Grand Total
		A-Not A: single	A-Not A: multiple	Ranking	
R-Indices calculation from pooled data	B	84.5	88.6	94.8	89.3
	C	84.3	88.2	94.0	88.8
	D	77.6	84.3	91.3	84.4
	E	76.6	80.5	90.5	82.5
	F	51.6	63.4	54.1	56.4
R-Indices calculated by averaging judges' R-Indices	B	84.9	92.2	96.1	91.1
	C	83.7	92.0	93.0	89.6
	D	75.4	90.0	90.5	85.3
	E	75.5	87.5	90.2	84.4
	F	51.1	68.3	54.8	58.1

From the table, it can be seen that the highest R-Indices tended to be obtained with the ranking protocol. It may be hypothesized that this was because of the forced-choice nature of ranking and the results concur with previous research where ranking was seen to confirm Brown's (1974) prediction that higher R-Indices would be obtained with ranking than with a rating procedure (O'Mahony et al., 1980; Ishii et al., 1992).

The R-Indices for the ‘A-Not A: multiple’ protocol were higher than those for the ‘A-Not A: single’ protocol. It may be hypothesized that this was because the presentation of multiple standards gave the judges a better idea of the concept defined by the sensory characteristics of ‘A’. Single presentation of ‘A’ would allow a concept to be formed, yet this concept could possibly be generalized so widely as to include some of the products from ‘B’ to ‘F’. Yet, in the ‘A-Not A: multiple’ protocol, where the products ‘B’ to ‘F’ were presented beforehand along with ‘A’, this would allow the judges to form separate concepts for all these products. This would control the generalization of the concept for product ‘A’. Thus, the boundaries of the concept for product ‘A’ would be better defined. This would lead to fewer errors in the A-Not A test. For the single protocol, where such boundaries were not well defined beforehand, the concept of product ‘A’ would need to be established during testing. This would result in a higher error rate. This concurs with previous research that multiple standards, giving examples of stimuli both within and outside a sensory concept, provide a better definition of the concept than merely giving a single standard (Ishii and O’Mahony, 1991).

Next, it is interesting to compare the mean of the R-Indices computed from individual judges with pooled R-Indices where data from all judges are entered into a single matrix. In the latter case, where data from different judges are pooled on to the same matrix, judges would have different criteria. This would cause what is known as boundary variance. Boundary variance is a concept used in scaling. It refers to the fact that judges space their numbers differently when they are making numerical estimates using rating scales. Another way of describing this is to say that the boundaries between the numbers vary among judges. For example, judges will not place the boundary between numbers 6 and 7 at the same level of intensity. Thus, this boundary varies among judges, resulting in boundary variance. In the same way, with the A-Not A test, the boundaries between the categories ‘sure’ and ‘not sure’ and between ‘not sure’ and ‘guessing’ vary among judges. This added boundary variance has the effect of depressing sensory indices of difference. Another way of considering boundary variance is that one person’s ‘sure’ is another person’s ‘not sure’ and so entering both their data into the same matrix can result in more artificial ties and reversals. In the case where R-Indices were computed from individual judges, a judge would be expected to keep his own criteria fairly constant during an experimental session. Thus his individual R-Index values would not suffer from boundary variance and not be depressed. Thus the mean of such values would be expected to be higher than pooled R-Index.

For the ranking protocol, indices computed from pooled data and averaged from individual data did not show any systematic variation (t-test, p=0.99). This is to be expected because for ranking the boundaries are fixed both within and between judges, consequently excluding boundary variance.

For the ‘A-Not A: multiple’ protocol, the effect of boundary variance in the pooled data is apparent. For all products, mean R-Indices were higher when calculated from individual judges (t-test, p=0.001). The same effect would be expected for the ‘A-Not A: single’ protocol but it was not apparent (t-test, p=0.15).

It would be difficult to argue that the lack of difference for the ‘A-Not A: single’ protocol was due to judges’ all assuming the same boundaries (sure vs not sure vs guessing) as with ranking. Instead, it may be hypothesized that with only the single presentation of product ‘A’ beforehand, it was difficult for individual judges to establish stable boundaries. Thus, whether R-Indices were calculated from pooled data or by averaging judges’ individual R-Indices, the boundaries would be unstable in both cases. Thus differences between the two computational methods would not appear and R-Indices would be depressed as seeing in Table 4.

It is worth returning to the hypothesis that the boundaries of the concept of product ‘A’ was better defined by prior presentation of all the products in the ‘A-Not A: multiple’ protocol. Such an argument implies a beta-criterion. However, if only product ‘A’ was presented beforehand, the judge may not be able to establish the boundaries of the concept (beta-criterion). He might be forced to use tau-criterion instead. If the test did not have sufficient replications, giving examples of ‘A’ and ‘Not A’, he would not gain enough conceptual information to establish beta-criterion. In this case, he would need to use a tau-criterion throughout testing. Thus, considering the limitation on number of replicates in sensory evaluation, the logical possibility exists that differences in the A-Not A protocols have the potential to induce different cognitive strategies. Furthermore, the A-Not A test has a commonality with the triangle, duo-trio, and same-different tests in that the attribute change is not specified; this later tests involve tau criteria. Only when the attribute is specified (2-AFC, 3-AFC), is the beta criteria involved.

Should the use of tau-criteria be the case that, an explanation for the difference between the ‘A-Not A: multiple’ and the ‘A-Not A: single’ protocols might be that for individual judges, tau-criteria are not as stable as beta-criteria. Therefore, the lack of stability of tau-criteria for individual judges in the ‘A-Not A: single’ protocol, would produce as much boundary variance as when data were pooled over judges.

Yet, more information is needed concerning the decision rules or cognitive strategies involved in the A-Not A test. It is not known at this point, whether slight differences in the instructions or procedure might elicit different cognitive strategies or only affect the perceptual learning process for establishing beta-criteria. It is also not known whether differences among judges in terms of their experience (prior familiarity) might not do the same. The latter is currently under investigation.

For products ‘B’ to ‘E’, R-Indices were higher for ranking than for the ‘A-Not A: multiple’ protocol than for the ‘A-Not A: single’ protocol. This was not the case for the product ‘F’ where R-Index values were close to 50% (chance level) except in the ‘A-Not A: multiple’ protocol. Obviously the difference between the product ‘F’ and ‘A’ was much smaller than other differences. It may be hypothesized that the reason that it was discriminated better by the ‘A-Not A: multiple’ protocol was that the ‘familiarization’ (prior presentation of ‘A’ and ‘Not A’ products) came closer to a warm-up procedure and thus elicited greater judge discriminability for such small difference (Dacrament, Sauvageot and Duyen, 2000; O’Mahony, Thieme and Goldstein, 1988; Thieme and O’Mahony, 1990).

The authors are aware that the ranking or categorization for similarity as in the A-Not A method does not use a univariate dimension; differences between products can be due to different attributes. As a tool, in sensory evaluation, such methods should be used with caution because rankings or categorization might depend on conceptual differences as well as sensory differences. For the present experiment, this was not an issue because comparisons were made using the same judges with the same idiosyncratic conceptualizations.

4.4 Conclusions

It was apparent from the data that the ranking elicited higher R-Index values than the A-Not A methods. It can thus be seen as more sensitive and useful replacement for the A-Not A methods, provided that a sensory panel is able to repeatedly retaste the products involved. Regarding the A-Not A methods, ‘familiarization’ (the prior presentation of all the test products) given to the judges before beginning the test seems to be important to stabilize the cognitive decision criteria (beta-criterion). It was hypothesized that when judges were not experienced with the test products enough to develop the concepts of all the test products, tau-criterion can also be used in A-Not A methods. For the R-Index, knowledge of the cognitive strategy is not necessary for its computation. However, differences in cognitive strategy can affect the level of performance and thus the R-Index. For example, an R-Index using a beta-criterion will be slightly higher than an R-Index using a tau-criterion (Noreen, 1981; Hautus and Irwin, 1995; Irwin and Francis, 1995).

5. COMPARISON OF PERFORMANCE IN THE A-NOT A, 2-AFC, AND SAME-DIFFERENT TESTS FOR THE FLAVOR DISCRIMINATION OF MARGARINES: THE EFFECT OF COGNITIVE DECISION STRATEGIES

H-S. Lee, D. van Hout, and M. J. Hautus

Published in: *Food Quality and Preference, 18*, 920-928 (2007)

Abstract

The performance of three different discrimination tests (A-Not A, 2-AFC, same-different) was investigated to explore the effects of varying aspects of the test protocols, such as the familiarization procedure and retasting of the reference (A), during testing on discriminability and the cognitive decision strategy used in the tests, when discriminating between the two different margarines. Seven judges, who were not familiar with margarine products, each gave 24 ratings for each of six protocols, resulting in 168 ratings in the pooled data, and from which R-Indices and d′ estimates were calculated. When both test products were presented beforehand for familiarization, judges adopted the beta cognitive decision strategy. When only the reference (A) was presented to the judges beforehand for familiarization, and the reference (A) was retasted before the test product either by prescription or at will, the tau cognitive decision strategy was adopted. When the number of samples tasted within a test increased, discriminability was considerably decreased. Such differences between test protocols were explained in terms of the concept formation of the test products, carry-over and fatigue effects, and memory problems caused by longer time-intervals between tastings.

5.1 Introduction

Sensory difference tests are a vital tool in the sensory evaluation of food. These measurements are important for quality control, determining the effects of ingredient change, processing change, or changes in packaging. They are also used in storage studies and product development studies involving product imitation (benchmarking). Common tests used in food science are the triangle, duo-trio, paired comparison (2-AFC), A-Not A, and same-different tests.

A-Not A, unlike the triangle and duo-trio tests, has no standard form. There are many different versions of the A-Not A test (Lawless and Heymann, 1996; Lee, van Hout and O'Mahony, 2006b). But in general, the term 'A-Not A test' applies to any procedure where the judge undertakes a familiarization phase to build a memory of the product (A) at the beginning of a testing session, and then is given a set of unknown products, some of which are 'A', and others of which are not. The task of the judge is to report whether each product presented was 'A' or 'Not A'.

For the familiarization phase, some versions of the test prescribe that only the product, 'A', is presented beforehand to the judges (Pfaffmann, Schlosberg and Cornsweet, 1954; Institute for Perception, 2003), while other versions of the test prescribe that both 'A' and 'Not A' products need to be presented beforehand for familiarization (ASTM, 1968; Meilgaard, Civille and Carr, 1991).

Versions of the A-Not A test can also differ in the way that they prescribe the use of a 'reminder'. In some versions of the test, during the test session the judge is given 'A' at various intervals, knowing its identity, as a reminder (Pfaffmann et al., 1954; Peryam, 1958), while in other versions of the test, no reminders are given (Meilgaard et al., 1991).

The protocol is usually applied to two products, one 'A' and another slightly different product to be discriminated from 'A' (Not A), yet it can also be used as a multiple A-Not A test, which includes one 'A' and many 'Not A' products (Meilgaard et al., 1991). Such multiple A-Not A tests were first introduced and used in food science by Mahoney, Stier and Crosby (1957), and Wiley, Briant, Fagerson, Murphy and Sabry (1957), and were later described as a standard method for R-Index computation (Brown, 1974; O'Mahony, 1979, 1986).

Same-different tests do not have a standard format either. There is a short version and long version of the same-different test. For the short version, a test is regarded as the presentation of just one of the four possible pairs of two products 'A' and 'B' (i.e., <AA>, <BB>, <AB>, or <BA>). The subject must state whether the two products presented in a test were the same or different (Avancini de Almeida, Cubero and O'Mahony, 1999; Cubero, Avancini de Almeida and O'Mahony, 1995; Delwiche and O'Mahony, 1996; Stillman and Irwin, 1995). For the long version, a test consists of the presentation of two pairs, one pair the same (<AA> or <BB>) and one pair different (<AB> or

<BA>). With this version, the judge is unaware that one pair is the same and the other different. As far as the judge is concerned, he responds to each pair as if he were performing two short version tests (Rousseau, Rogeaux and O'Mahony, 1999; Rousseau, Meyer and O'Mahony, 1998).

There is also a modified version of the short-version same-different test, in which the reference (the first product in a pair) is always the same product. This version is frequently called the 'single reference same-different test' or 'degree of difference test' (Rousseau et al., 1999; O'Mahony, 2005). It can be used for multiple product discrimination, and in this case it is generally called the 'degree of difference test' (The Institute for Perception, 2003). The degree of difference test was first developed for the discrimination of overall flavor differences between heterogeneous products (Aust, Gacula jr, Beard and Washam, 1985). In this case, the test was applied to multiple product discrimination using a six-point category scale, with the scale labels: 'No difference', 'Very slight difference', 'Slight difference', 'Moderate difference', Large difference', and 'Extremely large difference'.

When this modified version of the same-different test is used with two products, there are only two possible presentation pairs (<AA>, <AB>) unlike the four possible pairs that are used in the standard same-different test. While it is true that the judges are asked to respond whether the two products in a test are the 'same' or 'different', in the psychological literature, this is not recognized as a same-different test at all. This is because the information available from the products in the modified test is different from, and used differently to, the way information is used in the standard same-different test. Consequently, in the psychological literature this modified version of the same-different test is considered a member of a whole different class of tests, called reminder tasks (Macmillan and Creelman, 2005).

This reminder task can be considered an extended A-Not A test. For the A-Not A test, one of two products ('A' and 'Not A') is presented and the judge is asked to identify the product as 'A' or 'Not A'. Consider that a reminder (product A) is presented prior to each A-Not A test. Now a response of 'A' is equivalent to a response of 'same' and a response of 'Not A' is equivalent to a response of 'different'. Thus, this reminder task is the same as the A-Not A test, except that a reminder is given in every trial to refresh the memory of 'A'. We refer to this protocol as 'A-Not A reminder'.

Signal Detection Theory provides a theoretical approach to the measurement of sensitivity and bias (Green and Swets, 1988; Macmillan and Creelman, 2005). It has been applied in food science to provide a fundamental measure of difference, d′, which is the distance between the distributions of perceptual intensity of the two confusable food products. To calculate d′, it is important to know the cognitive process used in the test. For example, when the same judge performs both the 3-AFC and the triangle test with the same products, the judge will perform better on the 3-AFC than the triangle test because a different cognitive strategy for the two different tests is used, as Frijters first explained in 1979; that is, (later named) a skimming strategy (beta-criterion) for 3-AFC and a comparison of distance strategy (tau-criterion) for the triangle test (O'Mahony, 1995). This different performance will be manifest, for example, by different percentages correct for each test. Signal Detection Theory can take into account the effects of the specific structure of each test and the cognitive strategy adopted. Signal Detection Theory therefore provides a measure of performance, d′, that is the same for both tests. This measure represents the ability of the judge to discriminate between the two products, independently of the test used or the cognitive strategy adopted (Ennis, 1993; O'Mahony, Masuoka and Ishii, 1994; O'Mahony and Rousseau, 2002; Lee and O'Mahony, 2004).

To determine the cognitive process used in a test, it is necessary to first clearly define the test method. If the test method is not standardized, or is vague, it is difficult to create a Signal Detection Theory model for the test, and such a model is required to reveal the cognitive strategy that is used by the judge. Any modifications made to the test method may directly influence the cognitive process used in the test and require that a new model be developed. Additionally, without having a correctly developed model and consequently being able to determine the cognitive process used in the test, it is not possible to derive correct estimates of d′, or to study the sensitivity and power of the test method.

So the following questions arise: What cognitive strategies are involved in the A-Not A test? How would the differences across the various versions of the A-Not A test, such as different familiarization procedures, or the presentation of reminders, influence the cognitive strategy adopted?

The A-Not A test (without reminder) is equivalent to a standard Yes-No task in psychology. For a standard Yes-No task, in which a judge is presented with either 'A' or 'Not A' and must indicate which product was presented, the beta strategy is the only cognitive strategy that can be adopted (Macmillan and Creelman, 2005). For this case, d′ can be calculated from the standard Yes-No ROC curve (Green and Swets, 1988).

For the A-Not A with reminder test, there are two cognitive strategies that can be adopted: the tau decision strategy and the beta decision strategy. These two strategies have alternative names in the psychological literature. The tau strategy is called the 'differencing strategy' (Macmillan and Creelman, 2005) or the 'sensory difference decision rule' (Noreen, 1981), and the beta strategy is called the 'independent-observation decision rule' (Macmillan and Creelman, 1991) or the 'optimal decision strategy' (Irwin and Francis, 1995; Noreen, 1981) or the 'decisionally separable strategy' (Macmillan and Creelman, 2005).

These decision strategies have been investigated mainly in relation to the standard (short form) same-different test and it is in this context that we introduce these strategies. Most research that has investigated the cognitive strategy used in the same-different test has indicated the use of the tau-criterion (Irwin and Francis, 1995; Irwin, Stillman, Hautus and Huddleston, 1993; Rousseau, 2001; Rousseau et al., 1998; O'Mahony and Rousseau, 2002). A tau-criterion is a statement of how different two sensory experiences need to be to be reported as 'different'. It can be visualized as a sensory yardstick. If sensations, elicited by the two stimuli in the same-different test, are more different than the yardstick, the products will be reported as different; if not, they will be reported as the same.

The alternative decision strategy is to use a beta-criterion. In this case, the judge will hold two categories in his head, one corresponding to 'A' and the other to 'B'. The boundary between the two categories is the beta-criterion (Rousseau, 2001; Rousseau et al., 1998). For this strategy the judge will independently assign the two products presented in a same-different test to either category. If both products are assigned to the same category then the response will be 'same' otherwise the response will be 'different'. There is some evidence that judges can use the beta strategy for the same-different test (Lee, van Hout, Hautus and O'Mahony, 2007a; Irwin and Francis, 1995; Francis and Irwin, 1995).

These two decision strategies can also be applied to the A-Not A with reminder test. The models are given in detail by Macmillan and Creelman (2005). In short, the tau strategy requires a decision from the judge whether or not the perceptual difference between the reference and product samples exceeds their tau-criterion. The beta strategy, on the other hand, requires that the judge ignores the reference stimulus entirely, and instead makes a standard A-Not A judgment about the second product presented in the test.

The objective of the present study is to explore the results obtained from various test protocols and to use the observed differences as a basis for inference concerning the nature of the cognitive decision strategies that are used in various versions of the A-Not A test. The test protocols employed include two versions of the 2-AFC test (standard 2-AFC and 2-AFC combined with a reminder) and the same-different test. All protocols incorporated a rating procedure so that ROC curves could be generated.

5.2 Materials and Methods

5.2.1 Judges

Seven female panelists (age range, 35-39 yrs.), who were not experienced with margarine products, were tested. All were familiar with the A-Not A, 2-AFC, and same-different test methods.

5.2.2 Stimuli

Two commercial margarine products were obtained from the local supermarkets in Vlaardingen, Holland. These were: (A) Halvarine (Gouda's Glorie, Zeewold, NL), (Not A) Sunflower (Gouda's Glorie, Zeewold, NL). For the purposes of this article, these products will be referred to by their corresponding names 'A' and 'B'. All products were presented in 50-ml white plastic-lidded cups under red light to minimize any color and reflectance differences. To sample the product, judges removed the lid and used separate plastic teaspoons for each tasting. Products were served chilled (5°C) having been stored in a fridge until 5 min before serving. Between tastings, judges rinsed *ad-lib* with de-mineralized water. Before beginning each protocol, judges were allowed to eat Barber crackers (the Horizon Biscuit Company Ltd., England) if desired. After this all judges rinsed at least five times.

5.2.3 Procedure

Judges performed difference tests on the two margarines ('A' and 'B') using six separate protocols. The protocols studied are summarized in Table 1.

Table 5.1 defining characteristics of the investigated six protocols.

Protocol	**Familiarization**	**Retasting**	**Number of samples presented in each test**	**Number of responses required in each test**
1. A-Not A	Both reference (A) and test product (B)	No	One sample	One response
2. A-Not A voluntary reminder	Only reference product (A)	Yes, as often as needed	One sample	One response
3. A-Not A reminder	Only reference product (A)	Yes, in every trial	Two samples	One response
4. 2-AFC	Both reference (A) and test product (B)	No	Two samples	Two responses for each sample in a pair
5. 2-AFC reminder	Only reference product (A)	Yes, in every trial	Three samples	Two responses for each sample in a pair
6. Same-different	Both reference (A) and test product (B)	No	Two samples	One response for each pair

For protocol 1 (A-Not A), a version of a standard Yes-No task was used. For this protocol, two products ('A', 'B') were given beforehand for familiarization. Judges were able to taste the two products as much as desired until they felt they had become familiar with each product's sensory characteristic (at least 4 teaspoonfuls). Preliminary studies for margarine discrimination with judges not trained on the test products, and so unfamiliar with their sensory characteristics, indicated that such familiarization is necessary for the judges to establish decision criteria for the task, and to produce consistent results. They were then given six unknown samples, one by one, and required to report for each one whether it was 'A' or 'B' (Not A). For the six samples, three 'A' and three 'B' products were presented in random order. During testing, judges were not allowed to retaste any sample.

For protocol 2 (A-Not A voluntary reminder), a slightly different version of A-Not A was used. For this protocol, only ‘A’ was tasted beforehand for familiarization. Judges were able to taste ’A’ as much as desired until they felt they had become familiar with its sensory characteristics (at least 4 teaspoonfuls). They were then given six unknown samples, one by one, and required to report for each one whether it was ‘A’ or ‘B’ (Not A). During testing, ‘A’ could be retasted as often as desired.

For protocol 3 (A-Not A reminder), a version of a reminder task was used. As in protocol 2, only ‘A’ was tasted beforehand for familiarization; again, this could be done as often as desired. Judges were then given six unknown samples, one by one, and required to report for each one whether it was ‘A’ or ‘B’ (Not A). Additionally, during testing, ‘A’ was retasted at the beginning of every test before the test sample was tasted.

For protocol 4 (2-AFC), as for protocol 1, the two products (‘A’, ‘B’) were presented for familiarization as many times as required by the judges. Two products (‘A’, ‘B’) were presented in each test and judges were required to first identify which was ‘A’ and which was ‘B’ after having tasted both. They then gave sureness ratings for each sample. Three pairs (tests) were presented in a session and the presentation order of each pair was randomized. During testing, retasting was not allowed.

For protocol 5 (2-AFC reminder), a version of the 2-AFC test that includes a reminder was used. As for protocols 2 and 3, only ‘A’ was tasted beforehand for familiarization. The test procedure was the same as for protocol 4 except that ‘A’ was retasted at the beginning of every test before the two test samples were tasted.

For protocol 6 (same-different), a short version of same-different test was used. Over the four sessions, four possible pairs of products (<A, A>, <A, B>, <B, A>, and <B, B>) were each presented six times. As for protocols 1 and 4, two products (‘A’, ‘B’) were given beforehand for familiarization. In a session, six pairs (tests) were presented in random order and judges were required to give a response for each pair – ‘same’ or ‘different’ – resulting in the same number of ratings as for the other protocols.

For all protocols, except protocol 6, each binary response (‘A’ or ‘B’) was followed by a three-point sureness rating (sure/unsure/I don’t know but guess), resulting in six-point rating data. As a consequence of this design, each protocol (except protocol 6) resulted in 168 ratings for both ‘A’ and ‘B’ products in pooled data. For protocol 6, each response (‘same’ or ‘different’) for each pair was followed by a three-point sureness rating (sure/unsure/I don’t know but guess) resulting in 168 ratings for both ‘same’ and ‘different’ categories in pooled data.

Judges performed all six protocols in each 2.5-hour session, having 5 min breaks between protocols. There were four sessions, each spaced one-week apart. The order of presentation of the protocols was counterbalanced over sessions.

5.2.4 Data Analysis

The R-Indices and d′ estimates, which indicate the perceptual difference between the two margarine products, ‘A’ and ‘Not A’, were calculated for each protocol. R-Indices were computed in two ways: First, data from all judges were pooled into a single response matrix for each protocol, for each of which a single R-Index was computed (number of A or Not A = 168 = 7 judges x 24 tests). Second, R-Indices were computed individually for each judge within each protocol (number of A or Not A = 24 per judge). Mean R-Indices, across judges but within each protocol, were then calculated.

For difference testing, the R-Index represents a probability value, which can be loosely interpreted as the probability of distinguishing between the two products tested. For perfect discrimination, the probability of correct choice is 100% (R-Index = 1) while for chance discrimination it is 50% (R-Index = 0.5). Intermediate values indicate discrimination between chance and perfect; the greater the value, the greater the degree of discrimination between the two food products under test.

Strictly, the R-Index obtained from a standard A-Not A test (equivalent to standard Yes-No) is equivalent to the probability of choosing a reference product (A) over the other product (B) in 2-AFC. This is not true for R-Indices calculated using test methods other than standard A-Not A. Consequently, the R-Index obtained from different test methods (e.g., same-different, triangle, or A-Not A tests), measuring the discrimination of the same two products, will be different. Thus, if the degree of difference between two products needs to be compared over many experiments, the same measurement protocol should be used over those experiments.

Yet, the degree of difference between two products can be compared in terms of d′ estimates, even using the different measurement protocols, as long as the same judges are evaluating over the experiments. In order to do so, estimates of d′ were calculated from the pooled rating data, using both the beta model (beta strategy) and the tau model (tau strategy).

5.3 Results and Discussion

Table 5.2 presents the R-Index values computed in two ways. Brown (1974), when he developed his R-Index, wished for a measure that was free of assumptions. Unlike d′ values, the computation of the R-Index does not use assumptions to circumvent the differences between the experimental methods used to measure it.

Table 5.2 R-Index values (model-free statistic) calculated from rating responses using two ways of combining data from individual judges.

	Method of combining judge data	
	A	**B**
Protocol	**R-Indices calculated from pooled data**	**R-Indices calculated by averaging individual R-Indices**
1. A-Not A	0.85	0.81
2. A-Not A voluntary reminder	0.69	0.67
3. A-Not A reminder	0.75	0.76
4. 2-AFC	0.89	0.88
5. 2-AFC reminder	0.77	0.65
6. Same-different	0.69	0.67

Column A: R-Indices were computed between the two products, A and B (Not A), from pooled ratings over seven judges (‘B’ as signals and ‘A’ as noise).

Column B: R-Indices were computed between the two products, A and B (Not A), for each individual’s ratings (‘B’ as signals and ‘A’ as noise) and then averaged over the seven R-Indices.

Accordingly, the R-Index is prone to vary with experimental method. For example, an R-Index obtained from ranking will be higher than one obtained from rating (O’Mahony, Garske and Klapman, 1980; Ishii, Vié and O’Mahony, 1992; Lee et al., 2006b). Thus, the R-Index can be seen as a measure of ‘performance’ rather than a fundamental measure of difference. Because of this, the R-Indices in Table 2 should only be compared between protocols that used the same underlying test method.

As an example, considering the A-Not A and 2-AFC tests separately, the test protocol in which the familiarization procedure included both ‘A’ and ‘B’ (Not A) products resulted in higher R-Indices than the protocol in which a familiarization with only the reference (A) was used.

With respect to retasting the reference (A) product during the test in the A-Not A test, protocol 3 (A-Not A reminder) in which retasting of the reference (A) was mandatory on each test resulted in higher performance than protocol 2 (A-Not A voluntary reminder) in which the reference (A) could be retasted any time as often, or as little, as desired.

Comparing the two R-Index computation methods (Table 2), R-Indices calculated by averaging R-Indices obtained by individual judges were lower than R-Indices calculated from pooled data, except for protocol 3 (A-Not A reminder), indicating the possibility that judges were different in their sensitivity as well as the way that they used their decision criteria. This discrepancy between R-Indices derived using each of the two methods was most apparent in protocol 5 (2-AFC reminder).

Lee et al. (2007b) also compared protocols 1 and 2, but with multiple products rather than just two, and named the A-Not A test protocol in which a familiarization procedure with ‘A’ as well as many ‘Not A’ products, as ‘A-Not A: multiple’. The A-Not A protocol incorporating familiarization with only the reference (A) was named ‘A-Not A: single’. Lee et al. noted that ‘A-Not A: multiple’ resulted in higher performance than ‘A-Not A: single’, and the effect of boundary variance in the pooled data was apparent in ‘A-Not A: multiple’ but was not significant in ‘A-Not A: single’. They speculated that these differences occurred because the presentation of multiple standards in the familiarization phase for the ‘A-Not A: multiple’ gave the judges a better idea of the sensory characteristics of ‘A’, while the presentation of only a single product in the familiarization phase for the ‘A-Not A: single’ made it difficult for the judges to establish stable boundaries.

Lee et al. (2007) also hypothesized that this difficulty in establishing stable boundaries, because of insufficient familiarization in the ‘A-Not A: single’ protocol, may lead judges to use a tau-criterion. This would be of particular concern in the sensory evaluation of food, where replication is limited resulting in judges not gaining enough conceptual information throughout testing with which to establish a beta-criterion.

Thus, considering the differences between the sensory evaluation of foods and the psychophysics of visual or auditory stimuli, it is possible that different familiarization procedures (together with retasting schemes) in sensory evaluation promote different cognitive decision strategies, even for the standard A-Not A test. For example, a judge in protocol 1 (A-Not A) might tend to use the beta-criterion, but the judge might tend to use the tau-criterion in protocol 2. Even in protocol 1, it is possible that if judges were naïve to the test products and the familiarization phase was insufficient to establish a conceptual boundary between the two products (A and B) then the judges might use a tau-criterion by comparing the test product to the unstable memory of the reference product (A). This hypothesis still holds for the present experiment and will be tested by comparing various Signal Detection Theory models to the data to see which of these models correctly describe the relation between A-Not A and the other discrimination tests. The appropriate models will be indicated by estimates of d′ that are approximately the same across protocols.

As mentioned earlier, the standard A-Not A test, which is the same as the Yes-No test, provides a context in which only the beta strategy can be adopted in psychophysics (Macmillan and Creelman, 2005). For protocol 3 (A-Not A reminder), either the beta or the tau strategy can be used. When a beta strategy is adopted, performance in the A-Not A reminder test is the same as performance in the A-Not A test. However, when the tau strategy is used, performance in the A-Not A reminder test will be poorer by a factor of $\sqrt{2}$ than performance in the A-Not A (Yes-No) test (Macmillan and Creelman, 2005). For protocol 2, where the A-Not A reminder is voluntary, a tau strategy could also be adopted, at least in tests where the reminder was sampled. In fact, it could be argued that the judge voluntarily samples the reminder so that the test sample can be directly compared to it; that is, this protocol encourages the judge to use a tau strategy. Unfortunately, because protocol 2 cannot be modeled effectively, as there is no set procedure to model, it is hard to know what the effect of a tau strategy would be on performance, except to concede that performance will be worse than that obtained using a beta strategy.

For the 2-AFC test, both beta and tau strategies can be used; however these two distinct strategies result in the same performance (Macmillan and Creelman, 2005). This can be extended to the 2-AFC reminder method. Just like in the 2-AFC test, it is possible to use either the beta or the tau strategy in the 2-AFC reminder method. Performance in 2-AFC reminder will be the same regardless of which strategy is used. Furthermore, this performance will be the same as that in the 2-AFC test (Hautus, in preparation).

Performance in the 2AFC method is predicted to be better than that in the A-Not A (Yes-No) method by a factor of $\sqrt{2}$ (Green and Swets, 1988; Macmillan and Creelman, 2005).

To investigate these relationships for our data, estimates of d′ were calculated from the pooled rating data. Table 5.3 presents these estimates based on both the beta model (beta strategy) and the tau model (tau strategy).

Table 5.3 Estimates of *d'* derived using appropriate beta and tau models.

	A	B
	d′ estimates using beta model (variance of the d′ estimates)	d′estimates using tau model (variance of the d′ estimates)
1. A-Not A	**1.55** (0.017)	-
2. A-Not A voluntary reminder	**0.74** (0.014)	**1.05** (0.020)
3. A-Not A reminder	**1.01** (0.015)	**1.43** (0.021)
4. 2-AFC	**1.31** (0.019)	**1.31** (0.019)
5. 2-AFC reminder	**0.81** (0.015)	**0.81** (0.015)
6. Same-different	**1.40** (0.016)	**1.74** (0.009)

Column A: For protocols 1 to 5, ROC curves were fitted to the data in the 'Not A'/ 'A' matrices using maximum-likelihood estimation. For protocols 1, 4, and 5, the equal-variance model fitted better, while for protocols 2 and 3, the unequal variance model fitted better.

For protocols 4 and 5, a correction has been made according to the relation between 2-AFC and A-Not A:

d′ for 2-AFC= [z(H)-z(F)] / $\sqrt{2}$

For protocol 6, ROC curves were fitted to the data in the 'different'/ 'same' matrices using maximum-likelihood estimation in an ROC-fitting program for the same-different task assuming the beta strategy (optimal strategy) (Hautus, Irwin and Sutherland, 1994).

Column B: For protocols 2 and 3, the values in Column B were adjusted according to the hypothesized relation between A-Not A and A-Not A reminder:

d′ for A-Not A reminder = $\sqrt{2}$ [z(H)-z(F)]

For protocol 5, the same value as in Column A was calculated based on the hypothesis of Hautus, van Hout and Lee (in preparation) as below:

d′ for 2-AFC reminder = d′ for 2-AFC = [z(H)-z(F)] / $\sqrt{2}$

For protocol 6, ROC curves were fitted to the data in the 'different'/ 'same' matrices using maximum-likelihood estimation in an ROC-fitting program for the same-different task assuming the tau strategy (differencing strategy) (Hautus, 1992).

Observation of estimates of d′ in Table 3 indicates some interesting results. Estimates of *d'* for protocols 1 (A-Not A, 1.55) and 4 (2-AFC, 1.31) are not too dissimilar and provide a reference level of performance against which the other protocols can be compared.

Protocol 2 (A-Not A voluntary reminder) yielded low estimates of *d'* irrespective of the decision strategy assumed. This reinforces our previous comment that this protocol cannot be effectively modelled. It is clear, however, that the *d'* estimate from the tau model (d′ = 1.05) is the closest to, but considerably below, that for protocols 1 and 4.

For protocol 3 (A-Not A with compulsory reminder) a much clearer result is obtained. The *d'* estimate for the tau strategy (*d'* = 1.43) lies clearly between those for protocols 1 and 4. The *d'* estimate for the beta strategy (*d'* = 1.01) is clearly below that for the two reference protocols. This strongly suggests that the tau strategy was adopted for protocol 3, and strengthens the weaker finding that the tau strategy was also used in protocol 2.

Performance on protocol 5 (2-AFC with reminder) was surprisingly low (d′ = 0.81); much less than that obtained for protocol 4 for which the same performance is predicted. This suggests that this task was particularly difficult for these judges, perhaps because of the memory load required to compare two separate samples back to an earlier presented reference. This finding clearly requires further investigation.

Finally, protocol 6 (same-different) provided an estimate of *d'* of 1.40 when the beta strategy was assumed. This value is very close to that for protocol 3 (A-Not A reminder) with the tau strategy (*d'* = 1.43) and lies between the values obtained for the reference protocols (1 and 4). This strongly suggests that a beta strategy was adopted by judges in the same-different task.

Investigating this pattern of findings further reveals that, for untrained judges (and thus, unfamiliar with the product), familiarization with both 'A' and 'B' products prior to testing enabled the judges to differentiate between the sensory characteristics of 'A' and 'B', and thus establish a beta cognitive decision strategy, while familiarization with only the reference (A) promoted a tau cognitive decision strategy. This is an interesting and important result that requires further research.

Table 5.4 presents the ratios of the estimated values of *d'* for each protocol to that obtained for protocol 1 (A-Not A). Only the best ratios (closest to 1.0) obtainable from estimates of d′ in Table 3 are presented in Table 4.

Table 5.4 Comparison between protocol 1 (A-Not A) and the other five protocols expressed as a ratio of *d'* (protocol 1) to *d'* obtained in each of the remaining protocols (2-6).

Protocol				
2	**3**	**4**	**5**	**6**
A-Not A voluntary reminder	A-Not A reminder	2-AFC	2-AFC reminder	Same-Different
1.47	1.08	1.18	1.91	1.11

For A-Not A, 2-AFC and same-different, beta strategy was assumed.

For A-Not A voluntary reminder, A-Not A reminder and 2-AFC reminder, tau strategy was assumed.

It is immediately apparent from Table 4 that three of these ratios are close to 1.0, indicating that approximately the same value of *d'* was estimated for four of the protocols (1, 3, 4, and 6). This provides good evidence that the underlying Signal Detection Theory models used to estimate d′ are appropriate models for the data for the A-Not A, A-Not A reminder, 2-AFC, and same-different methods.

Macmillan and Creelman (2005) reviewed the previous studies in psychophysics comparing performance on Yes-No, 2-AFC and reminder paradigms, using visual or auditory stimuli (Creelman and Macmillan, 1979; Jesteadt and Bilger, 1974; Jesteadt and Slims, 1975; Vogels and Orban, 1986), and reported that 2-AFC results in the best performance and Yes-No in the worst, with the reminder task intermediate between them. However, as seen in Table 4, when the familiarization was implemented, the A-Not A task (which is equivalent to Yes-No) resulted in the best performance for flavor discrimination of the food stimulus, margarine. For flavor discrimination, some form of familiarization procedure is generally introduced before the test, because unlike visual or auditory stimuli, food stimuli are often complex and have large inherent variation. Thus, judges need to learn the natural variability of the sensory characteristics of the stimuli before they can develop appropriate conceptual boundaries. Rousseau et al. (1999) reported that such familiarization has an effect of stabilizing the decision criterion used for flavor discrimination of mustard. Therefore, the appropriate implementation of familiarization seems very important to the efficiency of the A-Not A test.

When food stimuli that do not have strong adaptation and irritant effects are compared, a warm-up procedure (Dacremont, Sauvageot and Duyen, 2000; O'Mahony, Thieme and Goldstein, 1988; Pfaffmann et al., 1954; Thieme and O'Mahony, 1990) is sometimes used to increase discrimination performance. Warm-up usually consists of tasting alternately and rapidly the two stimuli to be discriminated, allowing judges to focus on the differences between the two stimuli. This process reduces the dimensions of the perceptual space, and consequently increases discriminability (Ennis and Mullen, 1985). For the present study using margarine, the effect of a proper warm-up procedure was not investigated. This is because the repetitive alternate tasting of margarine samples produces relatively strong adaptation and fatiguing effects. Yet, the applied familiarization (exposure to both 'A' and 'B' products at least four times before the testing) in the A-Not A protocol appears to function as a warm-up procedure and improve performance.

The obtained ratios for protocols 2 (A-Not A voluntary reminder, ratio = 1.47) and 5 (2-AFC reminder, ratio = 1.91) are high and indicate a problem for the underlying models to account for these methods. This finding is consistent with our earlier analysis of these data. In addition to the effect of familiarization, described above, we now give further consideration to the possible causes of this outcome.

We have already indicated that, in the A-Not A voluntary reminder test (protocol 2), because retasting of the reference ('A') is not standardized, the protocol is not equivalent to the A-Not A reminder task (protocol 3). Consequently the data are fitted to the wrong model and inconsistent results are to be expected. Setting this major problem aside temporarily, a second reason for the poor performance of judges on this task could be because retasting was allowed any time as much as desired. In practice, the judges tasted the reference product many more times in protocol 2 than in the other protocols, before they made their judgment. This could result in significant carry-over, resulting in lower sensitivity in protocol 2. Although *ad-lib* mouth rinsing was instructed between each tasting, it is still possible that judges did not rinse as often as required to eliminate carry-over.

Thus, it is possible that carry-over had a similar effect, but to a lesser extent, for protocols 3 (A-Not A reminder), 4 (2-AFC), and 6 (same-different), each of which has two sample presentations, compared to one sample for A-Not A (see Table 4). The judges tended to rinse their mouths between tests, but not between the tastings within a test. So it is possible that, because of such physiological interference, the two-sample methods (protocols 3, 4, and 6) resulted in poorer sensitivity than the single stimulus method (protocol 1).

This reasoning can also be applied to the 2-AFC reminder test (protocol 5). As can be seen in Table 4, 2-AFC resulted in higher sensitivity than 2-AFC reminder. Familiarization with both products appears to have a stabilizing effect on decision criteria in 2-AFC resulting in better performance than achieved on the 2-AFC reminder test. Additionally, in 2-AFC reminder, there is one more sample to be compared in a test than in 2-AFC. Thus, besides the effect of familiarization and taste carry-over, or taste residuals, the interference and decay of memory caused by the complexity of the task and the longer time interval between tasting the samples, could be other reasons for the relatively low performance observed in the 2-AFC reminder test. This is consistent with previous reports that three-sample tasks are sometimes less sensitive to stimulus differences than two-sample tasks because of memory-related issues (Lau, O'Mahony and Rousseau, 2004).

In vision and audition, physiological interference from carry-over or fatigue is minimal, and the duration of a test is relatively short. In flavor discrimination, cognitive factors such as imperfect memory caused by longer test duration and more pervasive physiological effects (carry-over or fatigue) can play a considerable role, and thus decrease the performance of the judges. Consequently it is often necessary to place a practical limit on the number of samples that are presented in a session.

Standard Signal Detection Theory models assume that each sample is perceived independently from those that preceded (Green & Swets, 1988; Macmillan & Creelman, 2005). As described above, this assumption most certainly fails in flavor discrimination because of taste carry-over. This is particularly important within each test where mouth rinsing usually does not occur. Consequently it is necessary to explore extended Signal Detection Theory models to help account for this non-independence.

5.4 Conclusions

For flavor discrimination using unfamiliar judges, a 'familiarization' phase incorporating the presentation of the two products to be tested prior to testing seems to be important for judges to establish the beta cognitive decision strategy for standard A-Not A and 2-AFC tests. Given this type of familiarization procedure, judges also used a beta strategy for the same-different test. This familiarization also resulted in better performance. In the A-Not A reminder test, where familiarization included only the reference (A), and additionally the reference (A) was retasted before every test product, the tau cognitive decision strategy was adopted by the judges. The single-sample A-Not A method resulted in the highest sensitivity. Additionally, methods with two sample presentations (A-Not A reminder, 2-AFC, and same-different) without rinsing between tastings resulted in much higher sensitivity than the three sample method (2-AFC reminder) without rinsing between tastings. It was apparent from the data that unlike in vision and audition where carry-over, fatigue, and memory effects are limited, in flavor discrimination such physiological and cognitive interference could play a considerable role. Therefore, to predict sensitivity in flavor discrimination using Signal Detection Theory models, the test procedure should be carefully standardized in a way that such interfering factors are minimized. Also extended Signal Detection Theory models should be explored to take account for such complications in flavor discrimination.

6. INVESTIGATION OF TEST PERFORMANCE OVER REPEATED SESSIONS, USING SIGNAL DETECTION THEORY: 2-AFCR COMPARED TO A-NOT A AND 2-AFC

DANIELLE VAN HOUT, MICHAEL J. HAUTUS and HYE-SEONG LEE

Published in: *Journal of Sensory Studies, 26*, 311-321. (2011)

ABSTRACT

To investigate more flexible methods for measuring overall sensory differences, the performance of three attribute non-specified difference test methods was compared using Signal Detection Theory. A-Not A, 2-AFC, and 2-AFCR tests were performed with experienced subjects over repeated sessions. Learning effects were investigated to determine how much training would be needed before subjects could perform these tests consistently. Results in d´ showed that in early sessions the 2-AFC and 2-AFCR performed better (higher d´) than A-Not A, and that after two sessions the test performance stabilized on that level. Learning effects for the A-Not A test were poorer in the earlier sessions, with performance levels increasing, even after six sessions. In business situations where the tested products frequently vary, the use of 2-AFC and 2-AFCR would be recommended because they are more sensitive in the earlier sessions, requiring less training time than A-Not A.

PRACTICAL APPLICATIONS

Sensory difference tests are important tools for business decision-making. Food companies may save a lot of money by employing more effective methodologies. For sensory difference tests to be used accurately and efficiently, test performances need to be investigated with consideration for food discrimination constraints and objectives. This paper explored the performance of 2-AFC and 2-AFCR over several sessions in comparison to the results of a standard signal detection measure, A-Not A, to use the differences between these methods as a basis for inference concerning the cognitive strategies and relative test effectiveness of 2-AFC and 2-AFCR for food discrimination, in practical situations when both the reference product and test product(s) change frequently. This paper suggested the applicability of a non-specified 2-AFCR procedure for discriminating small differences without much prior training.

6.1 Introduction

Food industries use sensory tests to measure sensory properties of products in a qualitative or quantitative way. From the early stages of product development, all the way through to launch and post-launch of products on the market, sensory tests provide Research and Development project teams and marketers with important information on their products. One class of most commonly used sensory tests is the sensory difference test (Moskowitz *et al.* 2006). Difference tests are important tools for business decision-making and are often used during the various stages of product development. Difference tests are used to quantify the size of small differences between products that are considered to be confusable. When conducted with trained sensory panels, these tests can help a company understand the effects of changes in processes and ingredients, or capture sizes of differences between products in the marketplace. When conducted with consumers, these tests can reveal the degree of sensitivity of a group of consumers to product differences, or their ability to discriminate between products.

Many types of difference tests exist and selecting the most effective test method for a certain situation is often not an easy task. The effectiveness of difference tests can be examined in terms of three aspects:

- The first aspect is the aim of the test. Here we can distinguish between 1) measuring the sensory differences between products, often referred to as *analytical sensory evaluation* (Chae *et al.* 2010), and 2) measuring consumer perception of products, referred to as *consumer sensory evaluation* (O'Mahony and Rousseau 2002). With analytical sensory tests, it is important that results are consistent as tests are often compared across sessions. External factors should be controlled in such a way that they will not introduce additional noise in the data. Therefore these tests often take place under controlled conditions. With consumer sensory evaluation, external validity is important; tests should be predictive for the target group and should be as realistic as possible. This also means that the test should not require extensive training, as training would increase subject sensitivity and therefore make the results less predictive for reality; that is, less valid.
- The second aspect is the sensitivity and statistical power of the test. Tests should be powerful enough to pick up the relevant differences, so that a relatively low number of tastings is required. Yet, among the tests having similar power, one test can be more sensitive than the others due to better practical performance. The sensitivity of difference tests can be affected by details given in instructions, training and/or familiarization procedures preceding the tests, the number of products tested, and the sequences of tested products (García-Pérez and Alcalá-

Quintana, 2010; Lee and O'Mahony 2007; Lee *et al*. 2009; Kim *et al*. 2010). For example, 3-AFC is statistically more powerful than 2-AFC, but in many practical situations, 2-AFC is more sensitive than 3-AFC due to fewer stimuli (which minimizes forgetting), taste adaptation, and other sequence effects (Bi *et al*. 2010; Dessirier and O'Mahony 1999). This is what we call *applied power;* the applied power of 2-AFC is better than 3-AFC.

- The third aspect is the consistency of the results; also called reliability, which means that similar conclusions would be drawn if the test was repeated.

In this paper we will focus on *analytical sensory difference tests*, typically conducted with trained sensory panels. For analytical sensory tests the sensitivity and power are very important aspects of the tests, because more powerful methods require fewer tastings and therefore the testing will be less costly. The consistency of the tests is also important. When results from previous tests can be reliably compared to present ones, the cumulative knowledge of the effects of various factors, such as changes in product formulations, leads to a greater understanding of a product's sensory properties.

The effectiveness (superiority of sensitivity and power) of the most popular difference test types has been frequently studied, using approaches based on Signal Detection Theory and Thurstonian modeling. In the Thurstonian modeling context, to understand the test effectiveness and determine the applied power of the test it is necessary to investigate the performance of various difference test methods and determine which cognitive decision strategy subjects use in the test. Thus, the various difference tests used in food science can be classified into two groups, according to the task and the cognitive strategy used.

The first group contains the non-attribute specified overall difference tests, such as triangle, same-different, and duo-trio tests (Meilgaard *et al*. 1999). In the case of the triangle test, the task instructions are along the lines of "which product is the odd one" or "which product is *different* from the others". The commonly used cognitive decision strategy for this group is the comparison-of-distances strategy, which is sometimes called the tau strategy (Lee *et al*. 2007c; O'Mahony *et al*. 1994).

The second group contains tests that focus on certain attributes (e.g., '*saltiness*'). These tests are called attribute-specified or directional difference tests and examples are 2-AFC and 3-AFC tests (Meilgaard *et al*. 1999). In these tests, subjects are instructed, for example, "which sample is the *sweeter one?"* or "which sample is the *more intense one?"* (Reckmeyer *et al*. 2010). In food science, the assumed cognitive decision strategy for attribute specified 2-AFC tests is either the categorization or skimming strategy, which is sometimes called the beta strategy or the comparison of magnitudes strategy, respectively (Lee and O'Mahony 2004; O'Mahony *et al*. 1994).

Due to the type of cognitive strategies used for the different tasks, the attribute-specified difference tests are superior to the overall difference tests in terms of power. For example, when trying to detect a ‘Just Noticeable Difference’ of d-prime (d´) = 1, the required number of tastings varies considerably between test methods (Bi 2011; O’Mahony and Rousseau 2002). Ennis (1990) reports that for detecting a difference of d´=1, with type 1 error = 0.01, and power = 0.8, the number of tastings required for the overall difference tests was 366 (duo-trio), and 318 (triangle), yet the attribute-specified difference tests required only 34 (2-AFC), and 25 (3-AFC). This shows that the attribute-specified 2- and 3-AFC tests are considerably more powerful than the triangle and duo-trio tests, as they need ten times fewer tests to demonstrate significance; in the given example, about 30 instead of 300 tests. However, typically the problem with attribute-specified difference tests is that these can only be used when the type of difference between the products can be identified prior to the test (e.g., ‘sweetness’), and that subjects should understand the concept of this difference. In the majority of test situations this is not the case and non-attribute specified, overall difference tests are more frequently used. In practice within food companies, the triangle or duo-trio tests are often used as overall difference tests with a low number of tastings, which means that these difference tests are performed with very low power. This leads to the severe risk that important differences might be missed. As mentioned, the triangle test requires over 300 tastings to detect a threshold difference of d´=1. If only 30 tastings are conducted with this triangle test, it will only be powerful enough to detect a difference twice as large; that is, d´=2. The second problem with some conventional non-attribute specified tests like the triangle test is that they give unreliable results; many studies have shown triangle tests to be inconsistent over repetitions; this is because the increased exposure and learning of the products over repeated test sessions can influence test performance over different sessions by encouraging the cognitive strategy to shift from the comparison of distances strategy to the beta strategy (O’Mahony *et al*. 1994; O’Mahony 1995; Rousseau and O’Mahony 2000). The vulnerability of triangle tests to shifts in decision strategies make it difficult to model and obtain accurate sizes of differences, which complicates the integration of results from different test sessions. Therefore, alternative, more powerful non-attribute specified difference tests, should be explored in order to improve the efficiency of sensory testing.

One such alternative non-attribute specified difference test is the ‘warmed-up’ paired comparison (WU 2-AFC), which was reported to be more powerful than the triangle test (Thieme and O’Mahony 1990; Angulo *et al*. 2007). In this method, prior to the real test, subjects alternately taste the two products several times and thereby learn the nature of the difference between the two products. This warm-up procedure was introduced because no matter whether or not the attribute was specified for 2-AFC, subjects often did not perceive the difference between the two confusable products at first, but only started noticing it after tasting them a few times alternately (O'Mahony *et al.* 1988). Immediately after this warm-up procedure, subjects can perform the 2-AFC test(s) to identify which product is which. However, in situations in which more than one product needs to be compared to a reference product, such warm-up procedures become complicated, and when there are multiple ways in which the test products can differ from a reference product, it might not be feasible to use such an instant warm-up procedure. In quality control, for example, the objective is to make the subjects familiar with one fixed reference product. They learn to notice any type of deviation from that reference. The warm-up procedure will then lead to problems, because products to be compared with the reference could have sensory differences from the reference in various directions and this could confuse the subjects (Mata-Garcia *et al.* 2006).

Another non-specified difference test is the A-Not A test (Bi and Ennis 2001a,b; Pfaffmann *et al.* 1954; Peryam 1958). This test is procedurally identical to a nonadaptive form of the yes-no task in psychophysics (Hautus *et al.* 2010; Lee *et al.* 2007a) and is suitable for multiple types of products (Lee *et al.* 2007b). The A- Not A test is also useful for application in situations in which only one product can be tested at a time; for example, when testing products with slight differences in appearance that would be noticed when products are served side by side (Meilgaard *et al.* 1999). In the A-Not A test, subjects are familiarized with the reference product first and then they evaluate one product at a time and need to identify whether the product is the same as or different from the reference product. A sureness judgment is added to the task to get around the problem of response bias; that is, the fact that responses made by subjects are influenced by their internal criterion about when to call something different or the same (O'Mahony 1992). The crucial part of the A-Not A test is the familiarization of the subject to the reference product. When subjects are not sufficiently familiarized with the reference product, they will not be able to recognize the reference product in a blind test because of poor memory, resulting in low or no discrimination between the reference and test products. Thus for conducting A-Not A tests, discriminating between a fixed reference product and other test products, an appropriate familiarization procedure is essential (Kim *et al.* 2011). Yet, in industrial evaluation settings, where the test products and the reference products can change very frequently from session to session, subjects need to be able to update their memory to the new reference quickly, and having to conduct familiarization procedures for every set of evaluations is not a very efficient approach. A better way to introduce the total sensory perception of the reference product is needed; a way that will enable subjects to discriminate products in a difference test based on their overall sensory differences.

In recent studies (Lee *et al.* 2007a; Hautus *et al.* 2011a, 2011b), a 2-AFC with reminder test (2-AFCR) was compared to 2-AFC and A-Not A tests, and was proposed as an alternative non-attribute specified difference test. The 2-AFCR test is procedurally similar to a duo-trio test with a constant reference (Hautus *et al.* 2009), however the duo-trio test is normally associated with a specific set of requirements (familiarization, number of replications, etc) that are not required for the 2-AFCR test. In the 2-AFCR test, subjects first receive a sample of the reference product before tasting a pair of 2-AFC samples. After tasting the two test samples, subjects identify which of the two test samples is the reference product. In a recent study, the decision strategies used in the 2-AFCR test were investigated and it was confirmed that subjects could use either the tau or beta strategy, but that both strategies led to the same sensitivity, which is equal to the sensitivity of 2-AFC (Hautus *et al.* 2009). In the 2-AFCR test, memory of the reference product is constantly triggered by tasting the reminder product in each test and based on the memory of this reference product, subjects can choose the first or second sample in a pair in the same way as in attribute-specified 2-AFC tests. Therefore, 2-AFCR is considered to be a better alternative to the currently applied overall difference tests, being more powerful than triangle and duo-trio tests and thus more suitable for testing food discrimination.

The main contribution of this paper is to explore the potential of the 2-AFCR test for more flexible use in the food industry as a non-attribute specified analytical sensory difference test. To find out if this test is useful in situations where products to be tested are changed rapidly, the speed with which subjects can learn the sensory properties of the reference product, and their performance in the discrimination tests, needs to be examined. To explore the relative performance of 2-AFCR over repeated sessions, the sensitivity of 2-AFCR will be compared to 2-AFC and A-Not A over repeated tests.

- Theoretically the performance of 2-AFCR should be equivalent to 2-AFC (Hautus *et al.* 2009), but since three samples are involved in a 2-AFCR test, rather than two samples in 2-AFC, it is also possible that the *applied power* of 2-AFCR is lower than 2-AFC due to carry-over and fatigue effects (Lee *et al.* 2007a). On the other hand, although three samples are tasted in 2-AFCR, it is also possible that the aid of the reminder could make 2-AFCR superior to 2-AFC when familiarity to the reference is improved enough.
- When comparing 2-AFCR to A-Not A, it can be hypothesized that in early sessions 2-AFCR will perform better than A-Not A due to the aid of the reminder, but that in later sessions, when familiarity to the reference is sufficiently improved, A-Not A will outperform 2-AFCR because for A-Not A, memory of the reference would make the decision criterion more stable and therefore the test results more reliable. There would also be less carry-over effect for A-

Not A. This would mean that for 2-AFCR, less familiarization would be needed than for A-Not A, indicating that 2-AFCR would be a more flexible test for situations in which the reference product changes frequently.

There have been a few papers that studied the performance of various difference tests over repeated testing, but these papers mainly dealt with the way of data analysis (Hautus 1997; Lee and O'Mahony 2006). In the present paper for the three potential non-attribute specified, overall difference test protocols (A-Not A, 2-AFC, and 2-AFCR) the test performance and learning effects over repeated sessions are investigated to study ways to improve the effectiveness of analytical sensory panels in the food industry.

6.2 Materials and methods

6.2.1 Subjects

Seven experienced female panellists (age range, 40-62 years), who were not experienced with testing margarine type products, took part in the study. All were familiar with the A–Not A, 2-AFC, and 2-AFCR test methods.

6.2.2 Stimuli

Two commercial margarine products were obtained from the local supermarkets in Vlaardingen, Holland. In a preliminary study, a larger set of products were tested, these two products were found to be sufficiently confusable in overall sensory properties to be appropriate for this study. The selected products were: (A) Gouda's Glorie halvarine (Zeewolde, NL), and (B) Ruitjes halvarine (Hoogvliet supermarkets, NL). For the purposes of this article, these products will be referred to as 'A' and 'B'. All products were presented in 50 ml white plastic lidded cups under red light to minimize any colour and reflectance differences. To sample the product, subjects removed the lid and used separate plastic teaspoons for each tasting. For each product, one spoonful of product was tasted and swallowed. Products were served chilled (5 °C) having been stored in a fridge until 5 min before serving. Between tastings, subjects rinsed *ad-lib* with room temperature de-mineralized water (20-22°C). Before beginning each protocol, subjects were allowed to eat Barber crackers (the Horizon Biscuit Company Ltd., England) if desired. After this the subjects rinsed at least five times. In a preliminary study, the appropriate number of tastings to hold in a session, and an effective rinsing procedure were established to reduce effects of fatigue as much as possible.

6.2.3 Procedure

Subjects performed discrimination tests between the two margarines ('A' and 'B') using three separate protocols, which are summarized in Table 6.1.

Table 6.1 Defining characteristics of the three protocols investigated

Protocol	A-Not A	2-AFC	2-AFCR
Instruction for the task	Is the product the same or different from the reference?	Which product is A and which is B?	Which product is A and which is B?
Familiarization	Only reference product (A)	Only reference product (A)	None
Re-tasting of the reference as reminder	No	No	Yes, in every trial
Number of samples presented in each test	One sample	Two samples	Three samples
Number of responses required in each test	One response	Two responses	Two responses

Protocol 1 – 'A–Not A': A version of a standard Yes–No task was used (Macmillan and Creelman 2005). Product 'A', called the reference, was given beforehand for familiarization. Subjects were able to taste the reference product as much as desired until they felt they had become familiar with its sensory characteristics (at least 4 teaspoonfuls). In preliminary studies for margarine discrimination it was found that such familiarization is necessary for subjects to produce consistent results. Subjects who are not trained on the test products, and who are therefore unfamiliar with their sensory characteristics, need such familiarization to establish decision criteria for the task. They were then given twelve unknown samples, one by one, and needed to report for each one whether it was the same as 'A' or different from 'A' (Not A, here 'B') and how sure they were. Responses were given on a scale with six categories: "same as reference, sure", "same as reference, not sure", "don't know but guess it's same as reference", "don't know but guess it's different from reference", "different, not sure", "different, sure". For the twelve samples, six 'A' and six 'B' products were presented in random order to each subject. During testing, subjects were not allowed to re-taste any of the samples.

Protocol 2 – '2-AFC': As in the first protocol, product 'A' was called the reference and given beforehand for familiarization. Two products ('A' and 'B') were presented in each test and after tasting both samples, subjects were required to first identify which was the same as 'A' and which was different from ‘A’ (thus ‘B’) and then give sureness ratings for each sample, using a similar scale as in the first protocol. Six product pairs were presented in a session and the presentation order of each pair was randomized. During testing, re-tasting was not allowed.

Protocol 3 – ‘2-AFCR’: No product was given for familiarisation. The test procedure was the same as for the other protocols, except that ‘A’ was re-tasted every time before the two samples were tasted.

Subjects performed all three protocols in 8 sessions of 2.5 hours each, twice a week, each about 3 - 4 days apart. In preliminary experiments this was found to be the optimal number of tests, to minimise effects of fatigue in a session. Within a session, five minute breaks were taken between tests. To counterbalance order effects, including learning effects within a session, in odd-numbered sessions the tests were sequenced in the order 2-AFC, 2-AFCR, A-Not A; and in even-numbered sessions the tests were sequenced in the order A-Not A, 2-AFCR, 2-AFC.

6.2.4 Data Analysis

To compare the performance of the three protocols over repeated sessions, d′ estimates, which indicate the perceptual difference between the two margarine products, ‘A’ and ‘B’ (Not A), were calculated for each protocol by fitting receiver operating characteristic (ROC) curves. ROC-analysis software for the A-Not A task was used, based on maximum-likelihood estimation (software written by Hautus (1994), based on an algorithm by Dorfman and Alf (1969)). First, data from all subjects were pooled into a single response matrix for each session, and then the data for adjacent sessions (1st and 2nd, 3rd and 4th, 5th and 6th, and 7th and 8th) were combined to counterbalance order effects and to increase the response frequencies (Table 6.2).

Table 6.2 Response frequency matrices obtained over 7 judges for 4 sets of sessions. In signal detection analyses, we assigned A to be noise and B to be signal + noise

Protocol	**Session**		**Same as A**			**ifferent from A**			***N***
			sure	probably	guess	guess	probably	sure	
A- Not A	1-2	A (ref.)	6	23	3	10	28	14	*84*
		B	4	28	5	8	25	14	*84*
	3-4	A (ref.)	24	24	14	6	11	5	*84*
		B	1	7	9	4	21	42	*84*
	5-6	A (ref.)	43	17	0	1	14	9	*84*
		B	1	4	3	3	14	59	*84*
	7-8	A (ref.)	60	15	2	4	1	2	*84*
		B	1	0	0	0	2	81	*84*
2-AFC	1-2	A (ref.)	21	30	1	7	18	7	*84*
		B	5	20	7	3	23	26	*84*
	3-4	A (ref.)	53	18	2	3	6	2	*84*
		B	1	7	3	3	13	57	*84*
	5-6	A (ref.)	61	15	1	0	6	1	*84*
		B	1	4	2	2	15	60	*84*
	7-8	A (ref.)	71	7	2	0	4	0	*84*
		B	1	3	0	2	7	71	*84*
2-AFCR	1-2	A (ref.)	28	25	6	6	9	10	*84*
		B	8	10	7	6	25	28	*84*
	3-4	A (ref.)	47	20	8	4	3	2	*84*
		B	2	2	5	7	13	55	*84*
	5-6	A (ref.)	62	14	3	0	2	3	*84*
		B	3	2	0	1	5	73	*84*
	7-8	A (ref.)	77	4	1	0	2	0	*84*
		B	1	1	0	0	4	78	*84*

To determine the learning effects over the repeated sessions, a single d′ estimate was computed from the pooled data for each pair of two adjacent sessions (number of A or Not A = 84 = 7 subjects x 6 tests x 2 sessions).

To test whether the equal- or unequal-variance Signal Detection Theory model better fitted the pooled data obtained from the different sessions, the data were analyzed by separately fitting ROC curves based on each model (Table 6.3).

Table 6.3 Results of roc analyses for single stimulus judgments (β-strategy), applying the standard signal-detection model for A Not-A. Slope of the equal-variance model is 1.

Protocol	Session	Equal variance model (k=4*)			Unequal variance model (k=3)				GOF statistic ** (k=1)	
		d'	χ^2	p	d'_a	Slope (SE)	χ^2	p	LL difference	p
A-Not A	1-2	- 0.03	1.75	0.78	- 0.03	1.05 (0.15)	1.64	0.65	0.05	0.81
	3-4	1.54	0.73	0.95	1.54	1.01 (0.19)	0.73	0.87	0.00	0.99
	5-6	1.88	9.87	0.04	1.81	1.42 (0.36)	7.98	0.05	0.94	0.32
	7-8	3.68	12.37	0.01	3.49	0.40 (0.28)	3.84	0.28	1.16	0.27
2-AFC	1-2	0.82	7.34	0.11	0.82	1.00 (0.16)	7.33	0.06	0.00	0.99
	3-4	2.38	1.35	0.85	2.38	1.08 (0.29)	1.29	0.73	0.04	0.83
	5-6	2.78	2.80	0.59	2.78	1.09 (0.35)	2.72	0.44	0.03	0.85
	7-8	3.30	5.80	0.21	3.26	0.71 (0.32)	5.78	0.12	0.29	0.58
2-AFCR	1-2	0.95	4.22	0.38	0.95	1.09 (0.18)	3.97	0.27	0.12	0.72
	3-4	2.40	1.69	0.79	2.41	0.91 (0.23)	1.40	0.70	0.07	0.78
	5-6	2.80	6.26	0.18	2.74	0.59 (0.24)	4.29	0.23	0.93	0.32
	7-8	3.94	1.86	0.76	3.63	0.37 (0.32)	0.65	0.88	0.87	0.34

* degrees of freedom: 10 (5 points on ROC curve x 2 axes (horizontal + vertical)) – 7 (5 criteria to fit + intercept + slope) = 3; For unequal-variance model k=3, for equal-variance model k=4, as slope is fixed (slope = 1).

* * Chi-squared analysis of ROC Log Likelihood (LL) goodness-of-fit statistics between equal variance model and unequal variance model

To determine whether or not the additional parameter (slope) used in the unequal-variance model significantly improved the fit when compared to the equal-variance model, the difference between the maximized values of log likelihood (LL) from each of the two models was compared using the Chi-square distribution with one degree of freedom. The goodness-of-fit statistic, Chi-square (χ^2), and the probability that the data arose from the model given that the model was correct (p) were also noted. The criterion of p>0.001was used to indicate an acceptable model fit (Press, Teukolsky *et al.* 2007).

Another important parameter for describing the perceptual distributions is the slope of the linear ROC on inverse-normal (i.e., z-score) coordinates. This parameter is determined by the ratio of the variances of the two perceptual distributions that arise from the products. The equal-variance model has a slope of 1. The slope of the unequal-variance model can be either lower than 1, which means that the reference product has a larger variance than the other product, or higher than 1, which means the reference product has a smaller variance than the other product (Hautus *et al.* 2011a).

To calculate the true d´, the appropriate signal detection model for the test method should be applied. In the ROC-analysis software, all data were fitted using the standard model for A-Not A; therefore a correction needs to be made to the A-Not A d´ when data are not actually collected in an A-Not A test. The correction for data collected in 2-AFC and 2-AFCR tests is based on the relation in performance between 2-AFC/2-AFCR and A-Not A, which is 1/√2. Therefore d´ for 2-AFC and 2-AFCR is less than d´ for A-Not A by a factor of √2 (Hautus *et al.* 2009).

To determine whether the d´ values for repeated sessions or different protocols were significantly different, a Chi-Square (χ^2) test was conducted with null hypothesis: all are equal, and alternative hypothesis: one or more are different (as with ANOVA) (Bi *et al.* 1997). If the null hypothesis is rejected, additional analyses are needed to determine which products are different. Z-values, calculated for pairs of d´ values, estimate the probability of getting the observed level of difference in d´ when, in reality, there is no difference (Bi *et al.* 1997).

6.3 Results

The frequency distributions over the six response categories were noted for each product, 'A' and 'B' (Not A) for each of the three different protocols. The frequencies for each response category were pooled across all subjects and then across pairs of adjacent sessions to counterbalance order effects (see Table 2). For 2-AFC and 2-AFCR, it is usual to obtain one response from a comparison between two stimuli in a test pair. Yet, in the present study two responses were obtained; one response for each of the two stimuli in a test pair. Sureness ratings were then collected for each response. As shown in Table 6.2, this resulted in the same number of responses for 2-AFC and 2-AFCR as for the A-Not A protocol.

ROC curves were fitted to these data, considering each response for each product for all three protocols as single-stimulus judgments, using maximum-likelihood estimation based on the standard signal-detection model using the β-strategy. The results are shown in Table 3. The fit of the equal-variance model for the ROC curves was acceptable in all cases ($p \geq 0.01$) and the fit was not different from the fit of the unequal-variance model based on the log likelihood (LL) ratio fit ($p > 0.1$). Thus, it appeared that there is no reason to adopt the unequal-variance model rather than the equal-variance model. However, it is worth noting that in sessions 5-6 and 7-8 of the A-Not A protocol, the p-value obtained using the equal-variance model was pretty low, and in fact the unequal-variance model better fitted these data. Considering the fit of the unequal-variance model, the estimates of the slope for sessions 7-8 was smaller than 1, indicating a smaller perceptual variance for the reference product, while the slope for sessions 5-6 was greater than 1, indicating a larger perceptual variance for the reference product. However, the slope estimate for sessions 5-6 is within two standard errors of 1, indicating that it is not significantly different from 1. For sessions 7-8 this is almost also the case. Thus there is little evidence to support the unequal-variance model for these two pairs of sessions. The unequal-variance model also better fitted the data in sessions 5-6 and 7-8 for the 2-AFCR protocol. Unlike for the A-Not A protocol, here the slope for both sessions tended to be smaller than 1, indicating a smaller perceptual variance for the reference product (Hautus *et al.* 2008; O'Mahony and Hautus 2008). In this case, consideration of the standard errors of these two slope estimates shows that neither is significantly different from 1. This evidence taken collectively suggests that the equal-variance model is adequate to account for the data. This is important, especially for the 2-AFC and 2-AFCR data, as it means that the transformations used to calculate the corrected d′ estimates, discussed in the following paragraph, are legitimate for these data.

Table 6.4 presents the corrected d′ estimates and confidence intervals for each protocol in the four pairs of adjacent sessions as well as for all sessions pooled together.

Table 6.4 Transformed estimates of *d*' derived using the appropriate signal-detection models.

Session	1. A-Not A			2. 2-AFC			3. 2-AFCR		
	d′	95% Confidence Interval		*d′*	95% Confidence Interval		*d′*	95% Confidence Interval	
1-2	**-0.03**^{aA}	**-0.35**	**0.29**	0.58^{aB}	0.30	0.86	0.67^{aB}	0.39	0.95
3-4	1.54 ^{bA}	1.18	1.90	1.68 ^{bA}	1.27	1.99	1.70 ^{bA}	1.35	2.06
5-6	1.88 ^{bA}	1.49	2.27	1.97 ^{bA}	1.56	2.37	1.98 ^{bA}	1.57	2.39
7-8	3.68 ^{cB}	3.01	4.35	2.33 ^{bA}	1.86	2.81	2.79 ^{cA}	2.18	3.40
pooled	**1.51**	**1.33**	**1.69**	**1.52**	**1.38**	**1.66**	**1.59**	**1.44**	**1.74**

For *d*' computation, ROC curves were fitted to the data in the 'Not A'/ 'A' matrices using maximum-likelihood estimation using the equal-variance model.

For protocols 2 and 3, a correction has been made according to the relation between 2-AFC/2-AFCR and A-Not A: d′ for 2-AFC/2-AFCR= d′ for A-Not A/ $\sqrt{2}$

^{abc} *d*' estimates within the same column, not sharing the same letter are significantly different (chi-square test, p<0.05)

^{AB} *d*' estimates within the same row, not sharing the same letter are significantly different (chi-square test, p<0.05)

The performance of 2-AFC and 2-AFCR is predicted to be better than A-Not A by a factor of $\sqrt{2}$, so for these tests, d´ values and variances have been corrected accordingly (Green and Swets 1966; Hautus *et al*. 2009; Lee *et al*. 2007a). Results show an increase in d´ values over repeated sessions for all protocols, which means that for all protocols the discriminability between the two products increased over sessions, as subjects became more familiar with the reference (A) and the test product (B) (Figure 6.1).

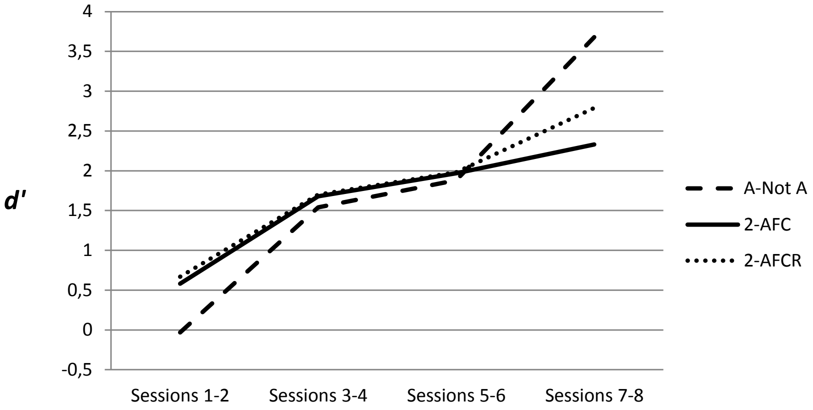

Figure 6.1 comparison of *d´* estimates for the three protocols across the four session pairs

As seen in Figure 6.1, the performance patterns for the different test types vary. In sessions 1-2, A-not A is the only test in which performance was not good enough to discriminate between samples, whereas with 2-AFC and 2-AFCR differences were detected (about d´= 0.6). In the middle sessions (3-4 and 5-6) all three tests performed similarly with d´= 1.5-2.0. In the last sessions (7-8), A-Not A showed a significant increase in discriminability to d´=3.68. 2-AFC showed no significant increase for the last sessions compared to the middle sessions (3-4 and 5-6) while 2-AFCR did show a significant increase over these same sessions.

Subsequent analyses were conducted to explore the effect of the specifically designed counterbalancing; that is, in half of the sessions A-Not A was tested first, in the other half 2-AFC was tested first. 2-AFCR was always tested in the middle. Data for tests sessions in which 2-AFC was tested first were pooled and compared to the pooled data for test sessions in which 2-AFC was tested last. The same analysis was conducted for the A-Not A tests. Results showed no significant difference in d´ values between first and last order in the session; for 2-AFC tests conducted first *d´* = 1.24 (1.04-1.44) and last *d´* = 1.62 (1.38-1.86), for A-Not A tests conducted first *d´* = 1.39 (1.14-1.64) and last *d´* = 1.62 (1.37-1.87). Although there were no significant differences between the same protocols run first and last in sessions, there was a tendency for performance to become better at the end of the session. This indicates that there are no problems with fatigue or loss of concentration during the session and confirms what was found when piloting the study.

6.4 Discussion

This study explores three difference test methods that can be used with analytical sensory panels, in situations where test products frequently change. Three non-attribute specified difference test protocols have been compared in terms of performance over repeated sessions; the A-Not A, 2-AFC, and 2-AFCR (2-AFC with reminder). Over the eight repeated sessions all three tests were conducted several times in a balanced order to study how quickly performance improves and at what stage sensitivity stabilizes.

It was found that the pattern of performance over repeated sessions was quite different for A-Not A, compared to 2-AFC and 2-AFCR. In the first sessions the A-Not A test was not able to pick up the difference, while the 2-AFC and 2-AFCR detected small but significant differences. In the middle sessions all tests detected medium size differences. The results of the last sessions for A-Not A show a ceiling effect. The difference is so large that subjects reach almost perfect performance, which makes it difficult to model accurately. For the last sessions of 2-AFC and 2-AFCR the *d´* values do not seem extremely large, but the response frequency distributions shown in Table 2 also indicate a ceiling effect for these tests as performance was close to perfect. The sensitivity of 2-AFC and 2-AFCR tests to small differences and sharp learning effects have resulted in such ceiling effects in sessions 7-8 even when *d´* values were only 0.58-0.67 in sessions 1-2. This again confirms that when the differences between products are expected to be relatively large (e.g., *d´*=2.5-3.0), 2-AFC and 2-AFCR tests should not be used because ceiling effects make the estimate of these relatively large d´ values less accurate. Instead, other difference tests could be used that are better at detecting larger differences without being affected by ceiling effects, like the same-different test (Rousseau *et al.* 1998). Alternatively, for measuring larger differences, ranking or descriptive analysis could also be used.

The specific counterbalancing of the tests in the sessions was a deliberate choice to limit the number of sessions required. Both A-Not A and 2-AFC were conducted first in half of the sessions and last in the other half; 2-AFCR was always tested in the middle. This could bias the test results as familiarity might increase over the course of a session. Subsequent data analysis did not show an order effect. No significant differences were found between results for sessions in which 2-AFC or A-Not A was the first or last test. As no order effects were detected between first and last, it is unlikely to expect any order effects for the middle positions and therefore it was thought that the performance comparisons among the three methods in the present experimental design were valid.

The results show a steep learning curve for all test methods over repeated sessions. Over the course of the sessions, subjects became more aware of the dimension of the difference between the products. The speed of this learning will be product specific; therefore it is important to conduct some preliminary experiments before starting tests on a new product type.

Based on the results from the present experiment, it can be concluded that 2-AFC and 2-AFCR tests could be used as overall difference tests and that they might have benefit over the A-Not A test in situations where products are changed frequently. As no specific familiarization is needed for 2-AFCR, this test could be considered as the more efficient form, providing better discrimination results with fewer repetitions. Thus, 2-AFCR is recommended for use as an analytical difference test in situations where products to be compared are changed regularly. The A-Not A test, although having the advantage of monadic assessment of multiple products, only performed well when subjects had sufficient familiarization with the reference product (Kim *et al.* 2011). Once the subjects have been familiarized, this test could be effective for testing small differences between multiple products and a reference product. The A-Not A test is therefore recommended for situations where a fixed reference product will be used over a period of time; for example, in quality control tests. To understand how much familiarization is required, good approaches are needed to monitor subject performance in difference tests. For sensory descriptive analysis there are various ways of monitoring panel performance, so that it can be determined whether the panel is sufficiently trained and ready for use (Meilgaard *et al.* 1999); for difference tests, however, there are no such approaches. One approach for the A-Not A test could be to examine the dispersion of response distributions over the response options of the reference product 'A'. If the proportion of "Not A" responses were high for the reference 'A', it may indicate that the subjects are suffering from response bias due to high within-product perceptual variations. A relatively low proportion of "Not A" responses for the reference 'A' is a good indication that subjects are familiarized enough with the reference product and that their criterion is stable. More investigation is needed to set accurate guidelines for this. In situations where there are constraints on training time, 2-AFCR is a safer test to use than A-Not A.

The present experiment is one of a series of experiments conducted to better understand how we can most effectively use difference tests in the food industry. In the sensory modalities of audition and vision, many psychophysical studies have been conducted in which cognitive decision strategies for various test methods have been studied in depth (e.g., Jesteadt and Bilger 1974; Hautus and Collins 2003; Hautus and Meng 2002; Hautus *et al.* 1994; Wang *et al.* 2005). It is important that the same depth of understanding also be reached in food science, taking product constraints into account.

However, physiological factors play an important role when testing food products, which makes it different from the studies in vision and audition. Examples of these physiological factors are carry-over effects; such as through adaptation from one product in the test to the next, which in turn causes order effects and sensory fatigue which can numb the senses or lower subject motivation. This limits the number of tests that can be conducted in one session (Lee, and O'Mahony, 2007) and thus tests requiring a lower number of tastings and that are less prone to order effects are much needed. In the present experiment the performance of three non-attribute specified sensory difference test methods were compared over repeated sessions to explore their performance; that is, their ability to detect differences. It was found that in early sessions both the 2-AFC and 2-AFCR performed well, while the A-Not A was not powerful enough to detect differences. In the middle sessions all tests were able to detect differences and in the last sessions ceiling effects occurred, which made it difficult to determine accurate sizes of differences. Based on these results, 2-AFC and 2-AFCR tests are recommended for flexible use in situations where test products frequently change. The A-Not A test is recommended for situations when products need to be assessed one-by-one; for example, when appearance differences will be noticed when products are compared side-by-side, or when a fixed reference product is used over a longer period of time. This test requires a longer training time to sufficiently familiarize subjects with the reference product.

7. GENERAL DISCUSSION

7.1 Summary of the main findings

The research in this thesis illustrates how Signal Detection Theory and Thurstonian modelling can be utilized in industrial sensory and consumer research to improve effectiveness of testing and quality of decision making based on sensory test results.

The review in Chapter 2 explained two ways in which we can use human subjects in sensory product tests during early stage product development (O'Mahony, 1995a). The first is what is called Sensory evaluation I (SE I), or *analytical sensory evaluation*, where subjects are used as instruments, sensors, as it were, to measure differences between products. The second is sensory evaluation II (SE II), or consumer sensory evaluation, where subjects are used as 'predictors' for a target population of consumers. There are different requirements for what makes a sensory test a good test; for SE I sensitivity and robustness is important and for SE II reliability and predictability. Signal Detection Theory and Thurstonian modelling provide a framework for integration of results from different sensory tests, which can be used to optimize test methods to meet the specific requirements. A range of sensory discrimination methods is compared in terms of effectiveness, and an approach is presented where the measure *d´* is used for comparing the sensitivities of naive consumers and trained sensory panels to small sensory differences. This comparison of sensitivities allows using the trained sensory panel to screen reformulated products to predict if consumers would notice the change, thereby saving costs and time.

The studies in Chapter 3-6 provide new insights on sensory difference tests for use in SE I and II. Chapter 3 demonstrated for the same-different test that subjects can learn to use an optimal decision strategy, if they had a warm-up task requiring the use of the optimal strategy, prior to the real test. It is expected that the positive effect of the warm-up procedure is stronger with real products and with other difference tests that require less samples, as Stocks, Van Hout and Hautus (2013, 2014) found in their research.

In Chapter 4 three multiple difference tests have been compared for use with margarines. Similarity ranking of all products compared to a reference product was found to be most sensitive, the A-Not A test with a voluntary reminder of the reference product was the least sensitive and the A-Not A with warm-up containing all products performed in-between the others. These differences in test performance demonstrate that prior exposure to the range of products in the test can improve test performance, as it helps subjects to get familiar with the differences that can be expected, leading to better concept formation, and more stable use of perceptual boundaries and cognitive strategies.

In Chapter 5 the performance of three discrimination tests (A-Not A, 2-AFC and same-different) was studied. The use of a reminder product of the reference during the trial lowered the performance, seemed to be caused by the extra product that needs to be tasted. In general it was found that the lower the number of samples in a trial, the more sensitive the method. Such differences in performance can be explained in terms of carry-over and fatigue effects, and memory problems caused by the time intervals between tastings. Again in this study it was found that familiarization with both test products prior to the trial improved the methods performance, as it allowed better concept formation of the products and elicits the use of the optimal cognitive strategy.

Chapter 6 provided insights on learning effects over the cause of several sessions of margarine testing, using three difference tests methods, A-Not A, 2-AFC and 2-AFCR without prior familiarization. For all tests, performance in the first sessions was low, and over the course of subsequent sessions performance increased and stabilized. This demonstrates the importance of familiarization before starting with the test. The two forced choice tests, 2-AFC and 2-AFCR, showed higher performance in earlier sessions, the A-Not A test showed steeper learning effects after a few sessions. These findings suggest that 2-AFC and 2-AFCR are tests that can be used more flexibly when different products are to be tested, or with different subjects without much prior training. Learning in the A-Not A test took more time but resulted in a higher test sensitivity in the later sessions. The A-Not A is therefore a very effective test to use with trained sensory panels in situations when a fixed reference product is used across studies. Another situation where the A-Not A could be effectively used is to measure the sensitivity of loyal consumers toward small product changes. Such consumers already have good representation of “their product” (the reference) in their memory, and can therefore perform this test without additional familiarization.

7.2 Limitations and directions for further research

Sensory product testing in a FMCG context has a number of limitations that need to be worked around to obtain accurate quantitative information that can be used for cumulating knowledge. The human subjects in the test provide noisy results as their responses are influenced by many internal and external variables. For methods to be effective, they should be designed in a way that reduces unnecessary, non-value adding, sources of variance. This can be achieved by using a task as straightforward as possible for the subjects, so that they can use the optimal cognitive strategy. In addition there is also a limit to the amount of testing a subject can do in a session. Therefore a method can only be effective if it is powerful enough to obtain robust results with a relatively low number of trials.

The research in this thesis demonstrates the potential of using Signal Detection Theory and Thurstonian modelling for improving the effectiveness of sensory testing in an industrial context, and of improving the quality of decision making based on sensory results. This work specifically focussed on the area of difference testing but it can be extended to adjacent areas such as preference, acceptance, and choice testing. This would extend knowledge on the meaningfulness of product differences, beyond only predicting what they could notice, towards whether they would accept it, or whether they would choose to buy the product again.

In the area of difference testing, ongoing research measures the effects of product complexity on the decision strategies used by subjects in a test and how this affects overall test performance (Stocks, Van Hout, and Hautus, 2013, 2014). Their research provides insights that can help improving the robustness of the difference tests for specific product types. Another area of further research is the investigation of more effective approaches for familiarization by framing the task for the subjects to induce a more holistic state of mind in the subjects. This approach is expected to increase the effectiveness and the predictive power of the test (Kim, Chae, Van Hout and Lee, 2011, 2014; Frandsen, Dijksterhuis, Brockhoff, Nielsen and Martens, 2007).

In sensometrics, the field of sensory data analysis and modelling, there is important progress made by integrating conventional statistical approaches for sensory data analysis with signal detection theory and Thurstonian modelling. By utilizing the strengths of both areas, novel approaches for data analysis will become available that allow better quantification of specific elements of the test, such as new modelling approaches for identifying differences between subjects and the effects of specific test design factors (Christensen, Cleaver, and Brockhoff, 2011; Christensen, Lee, and Brockhoff, 2012; Choi, Kim, Christensen, Van Hout, and Lee, 2014).

SUMMARY

Motivation

In the ‘fast moving consumer goods’ (FMCG) industry, results from sensory product tests form the basis for many important business decisions. Decisions on whether or not to launch new products, change existing products, or whether to continue with specific novel technological developments. It is therefore important that sensory tests are accurate and deliver robust results.

In most corporate functions where decision making takes place, ‘action standards’ are defined, based on direct measures of effect size. For example, in marketing whenever a new innovation is launched, post-launch marketplace performance measures are available, such as sales and complaints data, and in supply chain functions there is accurate information on logistics and pricing of raw materials, which makes it possible to make informed decisions.

For sensory product testing, direct measures of effect size are usually not possible, as these tests typically take place during the very early stages of development, long before the products enter the market. This means that we need to rely on indirect measures with human subjects, to predict potential in-market performance. The main problem in testing with human subjects is the noise in the measurement, caused by the various internal and external influences, both physiological and cognitive in nature that affects test performance.

The conventional statistical approach for sensory data analysis, also called deterministic models, or guessing models, are on their own not very well suited for modelling human data. The main reason is that statistical approaches do not take into account the physiological and cognitive factors in tests that can influence test performance. Consequently, sensory results such as the presence or absence of a difference between products depend on the method that is used. Due to this dependency, it is not possible to directly compare results from several sensory studies and build deeper understanding on effect sizes. Such understanding can be very valuable for a company as it will increase the predictability of the results, for example by calibrating differences ‘in-lab’ to ‘meaningful differences’ that consumers would notice when they use the product.

The problem of method-specificity of sensory results can be solved by using an alternative type of data analysis and modelling in addition to the traditional statistical approach. The framework of Signal Detection Theory (Green and Swets, 1966) and Thurstonian modelling (Thurstone, 1927) makes it possible to investigate the internal and external factors in tests and study how these factors influence subjects’ test performance. In particular, this framework can be used to study two key processes that take place in the human subject when performing a sensory test: the first is the *perceptual process* that integrates the information from the senses and the second is the *decision* *process* that uses the task instructions and the integrated information from the senses to base a decision on. Signal Detection Theory provides models of the relationships between different test methods, for example the differences in test sensitivity, and enables the development of tools for studying and optimizing test performance. It has led to a standardised measure of sensory difference, called d-prime (*d´*), available for most used test types. The *d´* allows to select the best test method for specific purposes and to compare results between different test types, e.g. comparing in-lab test results to findings of consumer in-home tests, so that optimal Action Standards can be defined.

Contribution

The research described in this thesis investigates how Signal Detection Theory and Thurstonian modelling can improve the effectiveness of sensory research in a FMCG industry, by exploring one specific test type of sensory tests; difference tests. Sensory difference tests are used to measure small differences between products, and can be used to answer important questions like: "*Are these two products similar in taste?*", "*Does this new ingredient make the product different?*", "*Will our consumers be able to notice the differences?*" The thesis focuses on two applications of Signal Detection Theory and Thurstonian Modelling. The first is to compare test methods, as there are many sensory difference tests available that largely differ in performance, and identify how to optimise the methods. With this knowledge, tailor-made methods can be designed for the specific test objectives and product types; tests that are more accurate and less costly. The second application of Signal Detection Theory and Thurstonian Modelling is to integrate results from different studies. This can improve the effectiveness of sensory testing in general, for example by relating sensory differences detected by a trained panel "*In Lab*" to differences found by consumers "*In Home*". Such knowledge can make future studies more predictive of what really matters to consumers, and improve the quality of decision making based on sensory results and reduce the overall number of testing that is required.

Summary of each chapter

Chapter 2 is a review of the current use of Signal Detection Theory and Thurstonian modelling for sensory science, and how it can be of benefit for the FMCG industry. It describes how different tests methods can be studied and optimised and requirements for what makes a good method for a specific objective. A distinction is made between two types of sensory product tests during early stage product development, introduced by O'Mahony (1995a). The first is what is called Sensory evaluation I (SE I), or *analytical sensory evaluation*, where subjects are used as instruments, sensors, as it were, to measure differences between products. The second is sensory evaluation II (SE II), or consumer sensory evaluation, where subjects are used as ‘predictors’ for a target population of consumers. There are different requirements for what makes a sensory test a good test; for SE I sensitivity and robustness is important and for SE II reliability and predictability. Signal Detection Theory and Thurstonian modelling provides a framework for integration of results from different sensory tests, which can be used to optimize test methods to meet the specific requirements. A range of sensory discrimination methods were compared in terms of effectiveness and an approach is presented where the measure *d´* is used for comparing the sensitivities of naive consumers and trained sensory panels to small sensory differences. This comparison of sensitivities allows using the trained sensory panel to screen reformulated products to predict if consumers would notice the change, thereby saving costs and time.

The studies in Chapter 3-6 provide new insights on sensory difference tests for use in SE I and II. In Chapter 3, we explore for the same-different test whether subjects can learn to use an optimal decision strategy. The same-different test is an alternative to the triangle and duo-trio tests and can have considerable more power if subjects use the optimal decision strategy. The main hypothesis in this chapter is that introduction of a 2-AFC-warm-up task, which is known to elicit an optimal decision strategy, would result in subjects continuing to use this strategy in the subsequent same-different test. The research was conducted with trained subjects using models systems and confirmed for some subjects that the optimal strategy in 2-AFC was carried over to the subsequent same-different test. The results indeed confirmed for the same-different test that subjects can learn to use an optimal decision strategy, if they had a warm-up task requiring the use of the optimal strategy, prior to the real test. It is expected that the positive effect of the warm-up procedure is stronger with real products and with other difference tests that require less samples, as Stocks, Van Hout and Hautus (2013, 2014) recently found in their research.

Using margarines, Chapter 4 compares three difference test methods for testing multiple products versus one reference product. The methods were (a) A-Not A with familiarization with the reference and voluntary reminder of reference during the test, (b) A-Not A with prior familiarization with all products without reminder of the reference during the test, and (c) similarity ranking of all products compared to the reference, have been performed by the same subjects. The results are compared in terms of R-indices. The test yielding the largest R-indices is the most sensitive in detecting differences. Similarity ranking of all products compared to a reference product was found to be most sensitive, the A-Not A test with a voluntary reminder of the reference product was the least sensitive and the A-Not A with warm-up containing all products performed in-between the others. These differences in test performance demonstrate that prior exposure to the range of products in the test can improve test performance, as it helps subjects to get familiar with the differences that can be expected, leading to better concept formation, and more stable use of perceptual boundaries and cognitive strategies.

Chapter 5 investigates the performance of three discrimination tests (A-Not A, 2-AFC and same-different) when discriminating between two different margarines. The effects of prior familiarization are investigated together with the effects giving a reminder of the reference product during the trials. Just like in chapter 4, in this study it was found that familiarization with both test products prior to the trial improved the methods performance, as it allowed better concept formation of the products and elicits the use of the optimal cognitive strategy. The use of a reminder product of the reference during the trial lowered the performance and seemed to be caused by the extra product in the trial. In general it was found that the lower the number of samples in a trial, the more sensitive the method. Differences in performance between test protocols were explained in terms of concept formation of the products, carry-over and fatigue effects, and memory problems caused by the time intervals between tastings.

Chapter 6 investigates more flexible methods for measuring overall sensory differences. The performance of three difference tests methods (A-Not A, 2-AFC and 2-AFCR) was compared in terms of learning effects over repeated sessions, again using margarines. Alls tests were conducted without prior familiarization, to determine how much training is needed before the subjects can perform the test consistently. For all tests, performance in the first sessions was low, and over the course of subsequent sessions performance increased and stabilized. This demonstrates the importance of familiarization before starting with a test. The two forced choice tests, 2-AFC and 2-AFCR, showed higher performance in earlier sessions, the A-Not A test showed steeper learning effects after a few sessions. These findings suggest that 2-AFC and 2-AFCR are tests that can be used more flexibly when different products are to be tested, or with different subjects without much prior training. Learning in the A-Not A test took more time but resulted in a higher test sensitivity in the later sessions. The A-Not A is therefore a very effective test to use with trained sensory panels in situations when a fixed reference product is used across studies. Another situation where the A-Not A could be effectively used is to measure the sensitivity of loyal consumers toward small product changes, such consumers already have good representation of “their product” (the reference) in their memory, and can therefore perform this test without additional familiarization.

Conclusions

Sensory evaluation plays an important role in the product development and improvement processes in the FMCG industry. During the different stages of such processes, various sensory and consumer test methods are used. Sensory test results are typically used on a test by test basis, because with the conventional statistical data analysis methods, results are method specific and therefore difficult to compare from one test to another.

With applications based on signal detection theory and Thurstonian modelling, results from different methods can be compared directly. This makes it possible to compare test results in an early stage of development with results from products when these are they are in the market to build strategic knowledge on what differences consumers would notice.

The research in this thesis demonstrates the potential of signal detection theory and Thurstonian modelling for improving the effectiveness of sensory testing in an industrial context, and of improving the quality of decision making based on sensory results. This work specifically focussed on the area of difference testing but it can be extended to adjacent areas such as preference, acceptance, and choice testing. This would extend knowledge on the meaningfulness of product differences, beyond only predicting what consumers could notice, towards whether they would accept it, or whether they would choose to buy the product again.

NEDERLANDSE SAMENVATTING

Motivatie

Sensorisch onderzoek speelt een belangrijke rol in de *Fast Moving Consumer Goods* (FMCG) industrie. Meestal in een vroeg stadium van productontwikkeling wordt er op basis van sensorische testresultaten besloten over het wel of niet doorgaan met de ontwikkeling van een nieuw of verbeterd product. Het is belangrijk dat dit soort beslissingen snel, maar ook weloverwogen, gemaakt worden op basis van betrouwbare informatie.

In de meeste afdelingen in een bedrijf waar beslissingen worden genomen over investeringen en voortgang van activiteiten, zoals inkoop, marketing en logistiek, zijn er directe effectmetingen van het resultaat mogelijk. Bij de introductie van een nieuw product heeft de marketing afdeling bijvoorbeeld al snel verkoopcijfers ter beschikking, en als een bestaand product veranderd wordt is er exacte informatie beschikbaar over de verandering in grondstoffen, productiekosten en opmerkingen en eventuele klachten van consumenten over het veranderde product. De sensorische kwaliteit van producten wordt vaak indirect gemeten. Lang voordat het product op de markt wordt gebracht, worden er sensorische testen uitgevoerd met proefpersonen om verschillen tussen producten te meten en idealiter het toekomstige succes van de producten op de markt te voorspellen.

Het feit dat sensorische testen uitgevoerd worden met proefpersonen brengt een aantal problemen met zich mee. Zo kan een proefpersoon in een sessie altijd maar een klein aantal producten testen, en is er een groot aantal interne en externe factoren van invloed op de manier waarop een proefpersoon de test uitvoert en een antwoord geeft. Deze factoren kunnen van cognitieve of psychologische aard zijn zoals het geheugen of de aandacht van de persoon, of van fysiologische aard zoals sensorische vermoeidheid of adaptatie. De effecten van deze factoren verschillen per test methode en zijn er de oorzaak van dat sommige testen gevoeliger en nauwkeuriger zijn dan andere.

Statistiek, de conventionele manier van sensorische data analyse, waarbij gebruik wordt gemaakt van statistische of deterministische modellen, vergelijkt de resultaten van een sensorische test met kansmodellen (denk aan het gooien met dobbelstenen). Daarbij wordt er geen rekening gehouden met de specifieke factoren die het functioneren van een test en de resultaten kunnen beïnvloeden. Hierdoor wordt beschikbare informatie genegeerd, en zal de conclusie of twee producten van elkaar verschillen sterk afhangen van de test methode die gebruikt wordt. Dit maakt het moeilijk om resultaten van verschillende testen met elkaar te vergelijken en op deze manier meer kennis over de producten op te bouwen.

Met Signaal Detectie Theorie en Thurstoniaanse modellen kunnen sensorische data op een andere manier geanalyseerd worden, als aanvulling op de statistische aanpak. Deze modellen maken het mogelijk om twee belangrijke processen te onderzoeken die plaatsvinden in de proefpersoon wanneer deze een sensorische test uitvoert. Het eerste is het *waarnemingsproces* waarin de informatie van de zintuigen wordt geïntegreerd en het tweede is het *beslissingsproces*, waarin de proefpersoon op basis van de zintuiglijke informatie en de instructies in de test een beslissing maakt en een antwoord geeft. Voor de meeste testmethoden zijn er modellen ontwikkeld afkomstig uit de Signaal Detectie Theorie. Deze modellen omvatten ook de onderlinge verbanden tussen methoden, bijvoorbeeld hoe ze verschillen in gevoeligheid. Hierdoor kunnen test methoden en hun resultaten direct met elkaar vergeleken worden, en verder verbeterd worden. Resultaten kunnen berekend worden in de eenheid d-prime (*d´*), een gestandaardiseerde maat van sensorisch verschil, die gebruikt kan worden voor vrijwel elke testmethode. Door te werken met *d´* kan een optimale testmethode voor een specifiek doel of product geselecteerd worden. Ook kan *d´* gebruikt worden om resultaten van verschillende testen te integreren, waardoor er strategische kennis opgebouwd kan over de productverschillen, bijvoorbeeld door verschillen waargenomen in een getraind sensorisch panel te vergelijken met verschillen die consumenten thuis waarnemen. Op deze manier kan beter bepaald worden welke productverschillen belangrijk zijn voor consumenten en kunnen er in een vroeg stadium in de productontwikkeling betere beslissingen genomen worden.

Bijdrage

In het onderzoek dat beschreven is in dit proefschrift, zijn verschillende toepassingen van Signaal Detectie Theorie en Thurstoniaanse modellen bestudeerd die de effectiviteit kunnen verbeteren van sensorische testen in de “Fast Moving Consumer Goods” industrie. Het richt zich op een specifieke soort sensorische testen, namelijk de sensorische verschiltesten. Sensorische verschiltesten kunnen kleine verschillen tussen producten meten, en worden gebruikt om belangrijke vragen te beantwoorden zoals: “*Smaken deze twee producten hetzelfde?*”, “*Verandert het product door deze verandering in ingredienten?*”, “*Kunnen onze consumenten het verschil opmerken?*”. Het onderzoek omvat twee verschillende toepassingen van Signaal Detectie Theorie en Thurstoniaanse modellen. De eerste toepassing is het vergelijken van verschiltestmethoden en het onderzoeken hoe deze verder verbeterd kunnen worden. Dit is belangrijk omdat er zo veel soorten verschiltesten beschikbaar zijn, die zeer kunnen verschillen in hun functioneren. Door kennis vergaard in dit onderzoek, kunnen bestaande methoden worden geoptimaliseerd voor specifieke doelstellingen en producten. De tweede toepassing van Signaal Detectie Theorie en Thurstoniaanse Modellen is het integreren van resultaten van verschillende studies. Dit kan de effectiviteit van het totale sensorisch onderzoek verbeteren, bijvoorbeeld door sensorische verschillen, gemeten door een getraind panel “*In*

Lab”, te vergelijken met verschillen waargenomen door consumenten “*In Huis*”. Door dit soort kennis kunnen toekomstige studies beter voorspellen welke aspecten werkelijk belangrijk zijn voor consumenten om een product te blijven kopen. Op deze manier kan de kwaliteit van de bedrijfsbeslissingen op basis van sensorische resultaten verbeterd worden, en tevens de benodigde hoeveelheid testen gereduceerd worden.

Belangrijkste bevindingen per hoofdstuk

Hoofdstuk 2 is een review van hoe Signaal Detectie Theorie and Thurstoniaanse modellen gebruikt zijn in de sensorische wetenschap, en geeft een visie van hoe dit de Fast Moving Consumer Goods industrie kan helpen beter gebruik te maken van sensorische onderzoeksresultaten. Verschillende test methoden worden vergeleken en er wordt beschreven hoe testen geoptimaliseerd kunnen worden voor bepaalde doeleinden. Er wordt een onderscheid gemaakt tussen twee soorten sensorisch onderzoek die gebruikt kunnen worden tijdens productontwikkeling, in navolging van O’Mahony (1995a). Het eerste type onderzoek wordt Sensory Evaluation I (SE I) genoemd, of *analytisch sensorisch onderzoek*, waarin mensen gebruikt worden als “instrumenten”, of “sensors”, om product verschillen te meten. Het tweede type onderzoek wordt Sensory Evaluation II (SE II) genoemd, of *sensorisch consumenten onderzoek*, waarin mensen gebruikt worden als “voorspellers” van hoe de consument in de doelgroep de producten waarneemt. De twee soorten sensorisch onderzoek stellen verschillende eisen aan testmethoden. Een SE I test moet gevoelig zijn zodat er maar een klein aantal metingen nodig zijn voor een goed resultaat. Bovendien dient een SE I test robuust te zijn zodat er bij herhaling van de test dezelfde conclusies worden getrokken. Een SE II test moet betrouwbaar zijn en valide (hij moet “meten wat je wilt meten”). Verder moet de test goed voorspellend zijn voor de doelgroep. Een aantal van de meest gebruikte sensorische verschiltesten zijn vergeleken in effectiviteit, en er worden aanbevelingen gegeven over welke testen het meest geschikt zijn voor de twee soorten sensorisch onderzoek. Een case-study met echte data illustreert hoe resultaten van verschillende studies geïntegreerd kunnen worden, zodat kennis opgebouwd kan worden over welke verschillen belangrijk zijn, en betere beslissingen genomen kunnen worden op basis van de resultaten. Met behulp van de maat d' is de gevoeligheid van getrainde sensorische panels vergeleken met de gevoeligheid van naïeve consumenten. Deze vergelijking van gevoeligheden voor kleine product verschillen maakt het mogelijk om een getraind panel te gebruiken om kleine productverschillen te meten en te voorspellen of de verschillen door consumenten worden waargenomen, waardoor er op tijd en kosten bespaard kan worden.

De studies in hoofdstukken 3-6 leveren nieuwe inzichten op, die gebruikt kunnen worden om sensorische verschiltesten effectiever te gebruiken voor SE I en II. In hoofdstuk 3 is voor de same-different test onderzocht of mensen een optimale beslissingsstrategie kunnen leren gebruiken. De same-different test is een alternatief voor de triangle en duo-trio test, en kan gevoeliger zijn als mensen de optimale beslissingsstrategie gebruiken. De hypothese in deze study was dat als de proefpersonen, voordat de echte test begon, een 2-AFC test met optimale beslissingsstrategie zouden uitvoeren, ze deze strategie ook zouden gebruiken in de daaropvolgende same-different test. Dit onderzoek is uitgevoerd met getrainde proefpersonen en zogenaamde modelsystemen van basissmaken opgelost in water. Dit onderzoek bevestigt dat sommige proefpersonen inderdaad tijdens een opwarming met een 2-AFC test, de optimale strategie van de 2-AFC overnamen in de daaropvolgende same-different test. Het kan worden aangenomen dat het positieve effect van een opwarmingsprocedure sterker is als het gaat om meer complexe echte producten, en ook bij verschiltesten die minder producten bevatten dan de same-different test, zoals Stock, Van Hout en Hautus (2013) recentelijk vonden in hun onderzoek.

Hoofdstuk 4 vergelijkt drie verschiltesten waarin meerdere margarines vergeleken worden met een referentie margarine. De onderzochte methoden waren (a) A-Not A met vooraf gewenning met de referentie en vrijwillige herinnering van de referentie tijdens de test, (b) A-Not A met vooraf gewenning met alle producten in de test maar zonder herinnering van de referentie tijdens de test en (c) ranking van alle producten op mate van verschil ten opzichte van de referentie. De testen werden allemaal uitgevoerd door dezelfde groep proefpersonen en de ranking test bleek het meest gevoelig te zijn, en de A-Not A met vrijwillige herinnering van de referentie was het minst gevoelig. De verschillen tussen de testen konden verklaard worden door de verschillen in testtypes, het vormen van een product concept in het geheugen, en de gebruikte beslissingstrategieën. Gewenning voorafgaand aan de test met de hele reeks producten hielp de proefpersonen om bekend te worden met de mogelijke verschillen in de test, waardoor ze zich de producten beter herinnerden en nauwkeuriger konden antwoorden.

In hoofdstuk 5 wordt het functioneren van drie verschiltesten onderzocht (A-Not A, 2-AFC en same-different) om de verschillen tussen twee margarines te onderzoeken. Effecten van gewenning voor de test, en het effect van het geven van een herinneringsproduct van de referentie werden onderzocht. Uit de resultaten bleek, net als in hoofdstuk 4, dat gewenning met beide producten voorafgaand aan de test, de testresultaten verbeterde. Dit kan verklaard worden doordat gewenning ervoor zorgt dat de proefpersonen de producten beter kunnen onthouden en een betere beslissingsstrategie kunnen gebruiken. Het gebruik van een herinneringsproduct van de referentie tijdens de test maakte de test minder gevoelig, wat waarschijnlijk veroorzaakt werd door het extra product dat geproefd moest worden. In het algemeen bleek dat hoe kleiner het aantal producten in een test, hoe gevoeliger de methode. Deze verschillen in functioneren van test methoden kunnen verklaard worden door overdrachtseffecten van de producten, vermoeidheidseffecten van de zintuigen en geheugeneffecten die veroorzaakt worden door de relatief lange tijdsintervallen tussen die producten die geproefd worden.

In hoofdstuk 6 is onderzocht in hoeverre verschiltesten flexibel zijn in gebruik, en van drie testmethoden (A-Not A, 2-AFC en 2-AFCR) zijn de leereffecten die optraden tijdens herhaalde sessies vergeleken, opnieuw met margarines. Dit geeft inzicht over de benodigde training voor de proefpersonen om de test consequent uit te kunnen voeren en tot robuuste resultaten te komen. De drie testen zijn uitgevoerd zonder enige vorm van gewenning met de producten vooraf. Uit de resultaten bleek dat alle testen tijdens de eerste sessie slecht functioneerden, en dat de resultaten verbeterden en stabiliseerden tijdens de latere sessies. Dit onderzoek benadrukt het belang van effectieve gewenningsprocedures voorafgaand aan een test, om de robuustheid van test resultaten te garanderen. De 2-AFC en 2-AFCR tests functioneerden beter tijdens de eerste sessies, terwijl de A-Not A test grotere leereffecten liet zien in de latere sessies. Deze bevindingen leiden tot de conclusie dat 2-AFC en 2-AFCR tests flexibeler zijn in gebruik. Met name wanneer de producten in studies van test tot test variëren, hebben deze testen weinig trainingstijd nodig. In de A-Not A test hadden de proefpersonen meer tijd nodig om de producten te leren. Maar wanneer dit eenmaal geleerd is, leidt het tot een gevoeligere test. De A-Not A test is daarom een zeer efficiënte test om te gebruiken met getrainde panels wanneer er bijvoorbeeld een vast referentie product gebruikt wordt in meerdere studies. Een andere situatie waar de A-Not A test gebruikt zou kunnen worden is om de gevoeligheid van trouwe consumenten ten aanzien van product verschillen te meten. Deze consumenten hebben een goede representatie van “hun product” in hun geheugen en kunnen hierdoor zonder veel training deze test uitvoeren.

Het laatste hoofdstuk, hoofdstuk 7 geeft een overzicht van de belangrijkste bevindingen van het onderzoek, en de aanbevelingen voor verder onderzoek.

Conclusie

Sensorisch onderzoek speelt een belangrijke rol tijdens productontwikkeling en productverbetering in de Fast Moving Consumer Goods industrie. Tijdens de opeenvolgende fasen van het productontwikkelingstraject wordt er gebruik gemaakt van verschillende sensorische testmethoden. De resultaten en conclusies worden meestal per study getrokken omdat met de conventionele statistische data analyse de resultaten van verschillende testmethoden niet rechtstreeks met elkaar vergeleken kunnen worden. Met behulp van toepassingen gebaseerd op Signaal Detectie Theorie en Thurstoniaanse modellen kunnen resultaten van verschillende testmethoden wel met elkaar vergeleken worden. Dit maakt het mogelijk dat sensorische testresultaten tijdens vroege stadia van productontwikkeling vergeleken kunnen worden met testresultaten tijdens en nadat het product op de markt is gebracht. Op deze manier kan er strategische kennis opgebouwd worden over welke productverschillen waarneembaar zijn voor consumenten tijdens het normale gebruik van het product.

Met behulp van Signaal Detectie Theorie en Thurstoniaanse modellen kan de effectiviteit van sensorisch onderzoek verbeteren, en kunnen er betere beslissingen gemaakt worden op basis van de nauwkeurigere resultaten. Dit onderzoek was speciaal gericht op verschiltesten, maar kan uitgebreid worden naar aanverwante gebieden zoals voorkeurstesten, acceptatietesten en keuzetaken, en vergroot de kennis over de mate waarin productverschillen belangrijk zijn voor consumenten. Niet alleen zou de vraag beantwoord kunnen worden of men de verschillen kan waarnemen, maar ook in hoeverre het product nog acceptabel is voor consumenten, en of men het product zou blijven kopen.

REFERENCES

A.S.T.M. (1968). *Manual on sensory testing methods stp 434*. American society for testing and materials, Philadelphia.

Angulo, O., Lee, H-S. & O'Mahony, M. (2007). Sensory difference tests: overdispersion and warm-up. *Food Quality and Preference*. *18*, 190-195.

Aust, L. B., Gacula J.R.M.C., Beard, S. A., & Washam, R. W. (1985). Degree of difference test method in sensory evaluation of heterogeneous product types. *Journal of Food Science, 50*, 511-513.

Avancini de Almeida, T.C., Cubero, E., & O'Mahony, M. (1999). Same-different discrimination tests with interstimulus delays up to one day. *Journal of Sensory Studies, 14*, 1-18.

Bartoshuk, L. M. (1968). Water taste in man. *Perception and Psychophysics, 3*, 69-72.

Bartoshuk, L. M. (1974). Nacl thresholds in man: thresholds for water taste or NaCl taste? *Journal of comparative and physiological psychology, 87*, 310-325.

Bartoshuk, L. M. (1978). The psychophysics of taste. *American Journal of Clinical Nutrition, 31,* 1068-1077.

Bi, J. (2005). Estimating population or group sensitivity and its precision from a set of individual d'. *British Journal of Mathematical and Statistical Psychology, 58*, 55-63.

Bi, J. (2011). Similarity tests using forced-choice methods in terms of Thurstonian discriminal distance, *d'*. *Journal of Sensory Studies, 26,* 151-157.

Bi, J., & Kuesten, C. (2012). Intraclass correlation coefficient (icc) – a framework for monitoring and assessing performance of trained sensory panels and panellists. *Journal of Sensory Studies, 27,* 352-364.

Bi, J., & Ennis, D.M. (2001a). Statistical models for the A-Not A method. *Journal of Sensory Studies, 16,* 215-237.

Bi, J., & Ennis, D.M. (2001b). The power of the A-Not A method. *Journal of Sensory Studies, 16,* 343-359.

Bi, J., & Ennis, D.M., O'Mahony, M. (1997). How to estimate and use the variance of *d*' from difference tests. *Journal of Sensory Studies, 12*, 87-104.

Bi, J., Lee, H-S. & O'Mahony, M. (2010). *d*' and variance of *d*' for four-alternative forced choice (4-AFC). *Journal of Sensory Studies 25*, 740–750.

Bi, J., Lee, H-S., O'Mahony, M. (2012). Statistical analysis of receiving operating characteristic (ROC) curves for the ratings of the A-Not A and the same-different methods. *Journal of Sensory Studies, 28*, in press

Boutrolle, I., Delarue, J., Köster, E.P., Aranz, D., & Danzart, M. (2009). Use of a test of perceived authenticity to trigger affective responses when testing food. *Food Quality and Preference, 20,* p418-426.

Brown, J. (1974) Recognition assessed by rating and ranking. *British Journal of Psychology, 65,* 13-22.

Casale, M., Ashby, F. G., & Standring, R. (2005). The role of perceptual learning in categorization. *Journal of Cognitive Neuroscience,* (Suppl. S.), 123-123.

Chae, J-E., Lee, Y-M., & Lee, H-S. (2010) Affective same-different discrimination tests for assessing consumer discriminability between milks with subtle differences. *Food Quality and Preference, 21*, 427-438.

Choi, Y-J., Kim, J-Y, Christensen, R.H.B., Van Hout, D.H.A. & Lee, H-S. (2013). Consumer discrimination tests: Superior performance of constant-saltier-reference DTF and DTFM to same-different tests for discriminating products varying sodium contents. *Food Quality and Preference*. Submitted.

Christensen, R. H .B., Cleaver, G. J., & Brockhoff, P.B. (2011). Statistical and Thurstonian models for the A-Not A protocol with and without sureness. *Food Quality and Preference, 22*, 542-549.

Christensen, R. H. B., Lee, H-S., & Brockhoff, P. B. (2012). Estimation of the Thurstonian model for the 2-AC protocol. *Food Quality and Preference, 24*, 114-128.

Clark-carter, D. (2003) Effect size: the missing piece of the jigsaw. *The Psychologist, 16,* 636-638.

Creelman, C. D. & Macmillan, N. A. (1979). Auditory phase and frequency discrimination: a comparison of nine paradigms. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 146-156.

Cubero, E., Avancini de Almeida, T. C., & O'Mahony, M. (1995). Cognitive aspects of difference testing: memory and interstimulus delay. *Journal of Sensory Studies, 10,* 307-324.

Dacremont, C., Sauvageot, F., & Duyen, T. H. (2000). Effect of assessor's expertise level on efficiency of warm-up for triangle tests. *Journal of Sensory Studies, 15*, 151-162.

Dai, H., Versfeld, N. J., & Green, D. M. (1996). The optimum decision rules in the same-different paradigm. *Perception and Psychophysics, 58*, 1-9.

Delwiche, J., & O'Mahony, M. (1996). Changes in secreted salivary sodium are sufficient to alter salt taste sensitivity: use of signal detection measures with continuous monitoring of the oral environment. *Physiology and Behaviour, 59,* 605-611.

Dessirier, J-M., & O'Mahony, M. (1999). Comparison of *d*' values for the 2-AFC (paired comparison) and 3-AFC discrimination methods: Thurstonian models, sequential sensitivity analysis and power. *Food Quality and Preference 10*, 51-58.

Dorfman, D. D. & Alf, E. A. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals – rating-method data. *Journal of Mathematical Psychology 6*, 487-496.

Ennis, D. M. (1990). Relative power of difference testing methods in sensory evaluation. *Food Technology, 44*, 114, 116, 117.

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies, 89,* 353-370.

Ennis, D. M. (2004). Personal communication.

Ennis, D. M., & O'Mahony, M. (1995). Probabilistic models for sequential taste effects in triadic choice. *Journal of Experimental Psychology, 21*, 1088-1097.

Ennis, D.M. (1998). Foundations of sensory science and a vision for the future. *Food Technology, 52,* 78-90.

Ennis, J.M. & Jesionka, V. (2011). The power of sensory discrimination methods revisited. *Journal of Sensory Studies, 26*, 371-382.

Francis M. A., & Irwin, R. J. (1995). Decision strategies and visual-field asymmetries in same-different judgments of word meaning. *Memory and Cognition, 23,* 301-312.

Frijters, J. E. R. (1982). Expanded tables for conversion of a proportion of correct responses (pc) to the measure of sensory difference (*d*') for the triangular method and the 3-alternative forced choice procedure. *Journal of Food Science, 47,* 139-143.

Frijters, J. E. R., Kooistra, A., & Vereijken, P. F. G. (1980). Tables of *d*' for the triangular method and the 3-afc signal detection procedure. *Perception and Psychophysiology, 27*, 176-178.

Frijters, J.E.R. (1979a). The paradox of discriminatory non-discriminators resolved.

Chemical Senses and Flavors, 4, 355-358.

Frijters, J.E.R. (1979b). Variations of the triangular method and the relationship of its uni-dimensional probabilistic models to three-alternative forced-choice signal detection theory models. *British Journal of Mathematical and Statistical Psychology, 32*, 229-241.

García-Pérez, M. A., and Alcalá-Quintana, R. (2010). The difference model with guessing explains interval bias in two-alternative forced-choice detection procedures. *Journal of Sensory Studies. 25*, 876–898.

Green, D.M., & Swets J.A. (1988). *Signal detection theory and psychophysics* (3rd ed.). Los Altos, California: Peninsula Publishing.

Halpern, B. P. (1986). What to control in studies of taste. In H. L. Meiselman & R. S. Rivlin (Eds.), *Clinical Measurement of Taste and Smell* (pp. 126-153). New York: Macmillan.

Hautus, M.J. (2012). SDT Assistant. (Version 1.0) [Software]. http://hautus.org

Hautus, M. J. (1994). Amplitude resolution by human and ideal observers for Rayleigh noise and other Gaussian processes. *PhD thesis, University of Auckland*. Http://hdl.handle.net/2292/2417.

Hautus, M. J. (1997). Calculating estimates of sensitivity from group data: pooled versus averaged estimators. *Behaviour Research Methods Instruments and Computers, 29*, 556-562.

Hautus, M. J. & Collins, S. (2003). An assessment of response bias for the same-different task: implications for the single-interval task. *Perception and Psychophysics, 65*, 844-860.

Hautus, M. J. & Meng, X. (2002). Decision strategies in the ABX (matching-to-sample) psychophysical task. *Perception and Psychophysics, 64,* 89-106.

Hautus, M. J., & Irwin, R. J. (1995). Two models for estimating the discriminability of foods and beverages. *Journal of Sensory Studies, 10*, 203-215.

Hautus, M. J., Irwin, R. J., & Sutherland, S. (1994). Relativity of judgements about sound amplitude and the asymmetry of the same-different ROC. *Quarterly Journal of Experimental Psychology, 47a,* 1035-1045.

Hautus, M.J. & Lee, A.J. (1998). The dispersion of estimates of sensitivity obtained from four psychophysical procedures: implications for experimental design. *Perception and Psychophysics, 60*, 638-649.

Hautus, M.J., O'Mahony, M., & Lee, H-S. (2008). Decision strategies determined from the shape of the same-different ROC curve: what are the effects of incorrect assumptions? *Journal of Sensory Studies, 23*, 743-764.

Hautus, M.J., Shepherd, D. & Peng, M. (2011a). Decision strategies for the A-Not A, 2-AFC and 2-AFC-reminder tasks: empirical tests. *Food Quality and Preference, 22*, 433-442.

Hautus, M.J., Shepherd, D. & Peng, M. (2011b). Decision strategies for the two-alternative forced choice reminder paradigm. Att. *Perception and Psychophysics 73*, 729-737.

Hautus, M.J., Stocks, M. A. & Shepherd, D. (2010). The single interval adjustment matrix (siam) yes-no task applied to the measurement of sucrose thresholds. *Journal of Sensory Studies, 25*, 940-955.

Hautus, M.J., Van Hout, D., & Lee, H-S. (2009). Variants of A-Not A and 2-AFC tests: signal detection models. *Food Quality and Preference, 20*, 222-229.

Irwin R. J., & Francis, M. A. (1995). Perception of simple and complex visual stimuli: decision strategies and hemispheric differences in same-different judgments.

Perception, 24, 787-809.

Irwin, R. J., & Hautus, M. J. (1997). Likelihood-ratio decision strategy for independent observations in the same-different task: an approximation to the detection-theoretic model. *Perception and Psychophysics, 59,* 313-316.

Irwin, R. J., Hautus, M. J., & Butcher, J. C. (1999). An area theorem for the same-different experiment. *Perception and Psychophysics, 61*, 766-769.

Irwin, R. J., Hautus, M. J., & Stillman, J. A. (1992). Use of the receiver operating characteristic in the study of taste perception. *Journal of Sensory Studies, 7*, 291-314.

Irwin, R. J., Hautus, M. J., & Francis, M. A. (2001). Indices of response bias in the same-different experiment. *Perception and Psychophysics, 63,* 1091-1100.

Irwin, R. J., Stillman, J. A., Hautus, M. J., & Huddleston, L. M. (1993). The measurement of taste discrimination with the same-different task: a detection-theory analysis. *Journal of Sensory Studies, 8,* 229-239.

Ishii, R., & O'Mahony, M. (1991). The use of multiple standards to define sensory characteristics for descriptive analysis: aspects of concept formation. *Journal of food science 56,* 838-842.

Ishii, R., Kawaguchi, H., O'Mahony, M., & Rousseau, B. (2007). Relating consumer and trained panels' discriminative sensitivities using vanilla flavored ice cream as a medium. *Food Quality and Preference, 18*, 89-96.

Ishii, R., Vié, A., & O'Mahony, M. (1992). Sensory difference testing: ranking R-indices are greater than rating R-indices. *Journal of Sensory Studies, 1*, 57-61.

Jesteadt W. (2005). The variance of *d*' estimates obtained in yes-no and two-interval forced choice procedures. *Perception and Psychophysics* 67, 72-80.

Jesteadt, W. & Bilger, R. C. (1974). Intensity and frequency discrimination in one- and two-interval paradigms. *Journal of the Acoustical Society of America, 55*, 1266-1276.

Jesteadt, W. & Sims, S. L. (1975). Decision processes in frequency discrimination. *Journal of the Acoustical Society of America, 57*, 1161-1168.

Kaplan, H. L., Macmillan, N. A., & Creelman, C. D. (1978). Tables of *d*' for variable-standard discrimination paradigms. *Behaviour Research Methods and Instruments, 10*, 796-813.

Kim H-J., Jeon, S-J., Kim K-O, & O'Mahony M. (2006). Thurstonian models and variance i: experimental confirmation of cognitive strategies for difference tests and effects of perceptual variance. *Journal of Sensory Studies, 21*, 465-484.

Kim, K. O., Ennis, D. M. & O'Mahony, M. (1998). A new approach to category scales of intensity ii. Use of *d*' values. *Journal of Sensory Studies, 13*, 251-267.

Kim, M-A., & Lee, H-S. (2012) investigation of operationally more powerful duo-trio test protocols: effects of different reference schemes. *Food Quality and Preference, 25*, 183-191.

Kim, M-A., Chae, J-E., Van Hout, D.H.A., & Lee, H-S. (2011). Discriminations of the A-Not A difference test improved when "A" was familiarized using a brand image. *Food Quality and Preference, 23*, 3-12.

Kim, M-A., Chae, J-E., Van Hout, D.H.A., & Lee, H-S. (2014). Higher performance of constant- reference duo-trio incorporating affective state of mind in comparison with balanced- reference triangle test. *Food Quality and Preference, 24*, submitted.

Kim, M-A., Lee, Y-M., & Lee, H-S. (2010). Comparison of *d*' estimates produced by three versions of a duo-trio test for discriminating tomato juices with varying salt concentrations: the effects of the number and position of the reference stimuli. *Food Quality and Preference, 21*, 504-511.

Lau, S., O'Mahony, M., & Rousseau, B. (2004). Are three-sample tasks less sensitive than two-sample tasks? Memory effects in the testing of taste discrimination. *Perception and Psychophysics, 66,* 464-474.

Lawless, H. T., & Heymann, H. (1996). *Sensory evaluation of food: principles and practices*. New York: Chapman & Hall.

Lee, H-S. & Kim, K-O., (2008). Difference test sensitivity: comparison of three versions of the duo–trio method requiring different memory schemes and taste sequences. *Food Quality and Preference, 19*, 97-102.

Lee, H-S., & O'Mahony, M. (2004). Sensory difference testing: Thurstonian models. *Food Science and Biotechnology, 13*, 841-847.

Lee, H-S. & O'Mahony, M. (2006). Sensory difference testing: the problem of overdispersion and the use of beta binomial statistical analysis. *Food Science & Biotechnology. 15,* 494-498.

Lee, H-S., & O'Mahony, M. (2007a). Difference test sensitivity: cognitive contrast effects. *Journal of Sensory Studies. 22*, 17-33.

Lee, H-S., & O'Mahony, M. (2007b). The evolution of a model: a review of Thurstonian and conditional stimulus effects on difference testing. *Food Quality and Preference. 18*, 369-383.

Lee, H-S., & Van Hout, D. (2009). Quantification of sensory and food quality: the r-index analysis. *Journal of Food Science, 74*, 57-64.

Lee, H-S., Van Hout, D.H.A., & Hautus, M. (2007b). Comparison of performance in the A-Not A, 2-AFC, and same-different tests for the flavour discrimination of margarines: the effect of cognitive decision strategies. *Food Quality and Preference, 18*, 920-928.

Lee, H-S., Van Hout, D.H.A., & O'Mahony, M. (2006). Sensory difference tests for margarine: a comparison of R-indices derived from ranking and A-Not A methods considering response bias and cognitive strategies. *Food Quality and Preference, 18*, 675-680.

Lee, H-S., Van Hout, D.H.A., Hautus, M.J., & O'Mahony, M. (2007a). Can the same-different test use a beta criterion as well as a tau criterion? *Food Quality and Preference, 18*, 605-613.

Lee, Y-M., Chae, J-E., & Lee, H-S. (2009). Effects of order of tasting in sensory difference tests using apple juice stimuli: development of a new model. *Journal of Food science, 74,* 268-275.

Macmillan, N.A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review, 84*, 452-471.

Macmillan, N.A. & Creelman C.D. (2005). *Detection theory: a user's guide (2nd ed.).* Cambridge university press, UK.

Macmillan, N.A., & Kaplan, H.L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98-1*, 185-199.

Mahoney, C. H., Stier, H. L. & Crosby, E. A. (1957). Evaluating flavour differences in canned foods. I. Genesis of the simplified procedure for making flavour difference tests. *Food Technology, 11*, 29-41.

Mata-Garcia, M., Angulo, O., & O'Mahony, M. (2006). On warm-up. *Journal of Sensory Studies 22*, 187-193.

Mcburney, D. C., & Pfaffmann, c. (1963). Gustatory adaptation to saliva and sodium chloride. *Journal of Experimental Psychology, 65,* 523-529.

Meilgaard, M., Civille, G., & Carr, T. (1999). *Sensory evaluation techniques, (3rd ed.).*

CRC press, US.

Moskowitz, H.R., Beckley J.H. and Resurreccion, A.V.A. (2006). Sensory and consumer research in food product design and development. Blackwell publishing, Ames, ia

Noreen, D. L. (1981). Optimal decision rules for some common psychophysical paradigms. In mathematical psychology and psychophysiology, vol 13: *proceedings of the symposium in applied mathematics of the American mathematical society and the society for industrial applied mathematics*. (s. Grossberg eds.) Pp 237-279, American Mathematical society, providence, r.i.

O'Mahony, M. (1972a). The interstimulus interval for taste: 1. The efficiency of expectoration and mouth rinsing in clearing the mouth of salt residuals. *Perception, 1*, 209-215.

O'Mahony, M. (1972b). The interstimulus interval for taste: 2. Salt taste sensitivity drift and the effects on intensity scaling and threshold measurement. *Perception, 1*, 217-222.

O'Mahony, M. (1972c). Salt taste sensitivity: a single detection approach. *Perception 1*, 439-464.

O'Mahony, M. (1979a). Salt taste adaptation: the psychophysical effects of adapting solutions and residual stimuli from prior tastings on the taste of sodium chloride. *Perception, 8*, 441-476.

O'Mahony, M. (1979b). Short-cut signal detection measures for sensory analysis. *Journal of Food Science, 44*, 302-303.

O'Mahony, M. (1986a). Sensory adaptation. *Journal of Sensory Studies, 1*, 237-257.

O'Mahony, M. (1986b). *Sensory Evaluation of Food*. Marcel Dekker, inc, New York.

O'Mahony, M. (1992). Understanding discrimination tests: a user-friendly treatment of response bias, rating and ranking R-index tests and their relationship to signal detection. *Journal of Sensory Studies, 7,* 1-47.

O'Mahony, M. (1995a). Sensory measurement in food science: fitting methods to goals. *Food Technology, 49,* p72-82.

O'Mahony, M. (1995b). Who told you the triangle test was simple? *Food Quality and Preference, 6,* 227-238.

O'Mahony, M. (2005). Personal communications.

O'Mahony, M., & Godman, L. (1974). The effect of interstimulus procedures on salt taste thresholds. *Perception and Psychophysics, 16*, 459-465.

O'Mahony, M., & Goldstein, L. (1987). Tasting successive salt and water stimuli: the roles of adaptation, variability in physical signal strength, learning, supra- and subadaptating signal detectability. *Chemical Senses, 12*, 425-436.

O'Mahony, M., & Goldstein, L.R. (1996). Effectiveness of sensory difference tests: sequential sensitivity analysis for liquid food stimuli. *Journal of Food Science, 51*, 1550-1553.

O'Mahony, M., & Hautus, M.J. (2008). The signal detection roc curve: some applications in food science. *Journal of Sensory Studies ,23*, 186-204.

O'Mahony, M., & Odbert, N. (1985). A comparison of sensory difference testing procedures: sequential sensitivity analysis and aspects of taste adaptation. *Journal of Food Science, 50,* 1055-1058.

O'Mahony, M., & Rousseau, B. (2002). Discrimination testing: a few ideas, old and new. *Food Quality and Preference, 14,* 157-164.

O'Mahony, M., Garske, S., & Klapman, K. (1980). Rating and ranking procedures for short-cut signal detection multiple difference tests. *Journal of Food Science*, *45*, 392-393.

O'Mahony, M., Masuoka, S., & Ishii, R. (1994). A theoretical note on difference tests:

models, paradoxes and cognitive strategies. *Journal of Sensory Studies 9*, 247-272.

O'Mahony, M., Thieme, U., & Goldstein, L. R. (1988). The warm-up effect as a measure of increasing the discriminability of sensory difference tests. *Journal of Food Science, 53*, 1848-1850.

Peryam, D. R. (1958) Sensory difference tests. *Food Technology, 12*, 231-236.

Pfaffmann C., Schlosberg, H. & Cornsweet, J (1954). Variables affecting difference tests. *In food acceptance testing methodology* (ed. Peryam, D. R., Pilgrim, F. J. & Peterson, M. S.) Quartermaster food and container institute, Chicago. (pp. 4-20).

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2007). *Numerical recipes*. *The art of scientific computing*. (3rd ed). Cambridge university press, London, UK

Reckmeyer, N. M., Vickers, Z. M. & Casallany, A.S. (2010). Effect of free fatty acids on sweet, salty, sour and umami tastes. *Journal of Sensory Studies, 25,* 751-760.

Rousseau, B. (2001). The b-strategy: an alternative and powerful cognitive strategy when performing sensory discrimination tests. *Journal of Sensory Studies, 16*, 301-318.

Rousseau, B., & O'Mahony, M. (2000). Investigation of the effect of within-trial retasting and comparison of the dual-pair, same-different and triangle paradigms. *Food Quality and Preference, 11*, 457-464.

Rousseau, B., & O'Mahony, M. (2001). Investigation of the dual-pair method as a possible alternative to the triangle and same-different tests. *Journal of Sensory Studies, 16*, 161-178.

Rousseau, B., Meyer, A., & O'Mahony, M. (1998). Power and sensitivity of the same-different test: comparison with triangle and duo-trio methods. *Journal of Sensory Studies 13,* 149-173.

Rousseau, B., Rogeaux, M., & O'Mahony, M. (1999). Mustard discrimination by same-different and triangle tests: aspects of irritation, memory and t criteria. *Food Quality and Preference, 10*, 173-184.

Rousseau, B., Stroh, S., & O'Mahony, M. (2002). Investigating more powerful discrimination tests with consumers: effects of memory and response bias. *Food Quality and Preference, 13*, 39-45.

Santosa, M., & O'Mahony, M. (2008). Sequential sensitivity analysis for same-different tests: some further insights. *Journal of Sensory Studies, 23*, 267-283.

Santosa, M., Hautus, M.J., & O'Mahony, M. (2011). ROC curve analysis to determine effects of repetition on the criteria for same-different and A-Not A tests. *Food Quality and Preference, 22,* 66-77.

Sepulveda, D.R., Chacon, R., Clarck, S., Olivas, G.I., & Jimenez, J. (2011). influence of chewing gum on the discrimination efficiency of 2-afc sensory tests. *Journal of Sensory Studies, 26*, 401-408.

Sorkin, R. D. (1962). Extensions of the theory of signal detectability to matching procedures in psychoacoustics. *Journal of the Acoustical Society of America*, *34*, 1745-1751.

Stillman, J. A., & Irwin, R. J. (1995). Advantages of the same-different method over the triangular method for the measurement of taste discrimination. *Journal of Sensory Studies, 10*, 261-272.

Stocks, M.A., Van Hout, D.H.A., & Hautus, M.J. (2013). Cognitive decision strategies adopted by trained judges when discriminating aqueous solutions differing in the concentration of citric acid. *Journal of Sensory Studies, 28,* 217-229.

Stocks, M.A., Van Hout, D.H.A., and Hautus, M.J. (2014). Cognitive decision strategies adopted by trained judges when discriminating food products differing in the concentration of sucrose and citric acid. *Food Quality and Preference,* submitted.

Tedja, S., Nonaka, R., Ennis, D. M., & O'Mahony, M. (1994). Triadic discrimination testing: refinement of Thurstonian and sequential sensitivity analysis approaches. *Chemical Senses, 19*, 279-301.

The Institute for Perception (2003). Ifprograms user's manual. Pp. 27-28. The institute for perception, Richmond, VA.

Thieme, U., & O'Mahony M. (1990). Modifications to sensory difference test protocols: the warmed-up paired comparison, the single standard duo-trio and the a-not a test modified for response bias. *Journal of Sensory Studies, 5*, 159-176.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review 34*, 273-286.

Van Hout, D., Hautus, M. & Lee, H-S. (2011). Investigation of test performance over repeated sessions using signal detection theory: comparison of three nonattribute-specified difference tests 2-AFCR, A-Not A and 2-AFC. *Journal of Sensory Studies, 26*, 311-321.

Van Hout, D., Dijksterhuis, G.B., & Groenen, P.J.F. (2014). From sensory panels to consumers; utilizing signal detection theory to increase the effectiveness of sensory evaluation. *Journal of the Science of Food and Agriculture*, submitted.

Vié, A., & O'Mahony, M. (1989). Triangular difference testing: refinements to sequential sensitivity analysis for predictions for individual triads. *Journal of Sensory Studies, 4,* 87-103.

Vogels, R. & Orban, G. A. (1986). Decision processes in visual discrimination of line orientation. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 115-143.

Wang, M., Irwin, R. J. & Hautus, M. J. (2005). Detection-theoretic analysis of same-different judgments for the amplitude discrimination of acoustic sinusoids. *Journal of the Acoustic Society of America, 117*, 1305-1313.

Wichchukit, S., & O'Mahony, M (2010). A transfer of technology from engineering: use of roc curves from signal detection theory to investigate information processing in the brain during sensory difference testing. *Journal of Food Science, 75*, 183-193

Wiley, R.C., Briant, A.M., Fagerson, I.S., Murphy, E.F. & Sabry, J.H. (1957). The northeast regional approach to collaborative panel testing. *Food Technology, 11*, 43-49.

Wolf-Frandsen, L., Dijksterhuis, G.B., Brockhoff, P.B., Holm Nielsen, J., & Martens, M. (2007). Feelings as a basis for discrimination: comparison of a modified authenticity test with the same-different test for slightly different types of milk. *Food Quality and Preference, 18*, 97-105.

Wong, D. (1997). Cognitive strategies of the triangle difference test. *MSc thesis, university of California, Davis*.

ABOUT THE AUTHOR

Danielle van Hout was born on the 31st of May in 1972 in Kerkrade, The Netherlands. After completing her secondary school at the Sancta Maria College in Kerkrade, she studied Food and Business at the Hogeschool Zuyd, in Heerlen. This study aimed to deliver Food marketers with broad skills in the field of Food Science, Marketing and Economics. During this study she conducted two internships on sensory evaluation topics. The first was to set-up descriptive sensory panel for Verkade Chocolates, and the second to implement a sensory quality system at Campina – Melkunie, Mona. After graduation in 1994, Danielle started working at Unilever Research and Development in Vlaardingen as a sensory panel leader. Through the years she has been taken on various different roles in the research organisation; sensory team leader, global innovation project leader, expertise team leader, and her current role is that of science leader in the global Strategic Science group. During her 18 years career in Unilever she studied and implemented many different sensory and consumer tests for the various research and innovation projects. In particular in the area of signal detection theory and Thurstonian modelling she collaborates with various leading academic scientists, researching more effective sensory and consumer test methods. Many of the results are published in peer-review journals.

PUBLICATIONS

List of authors peer-reviewed publications to date

Choi, Y-J., Kim, J-Y, Christensen, R.H.B., Van Hout, D.H.A. & Lee, H-S. (2014). Consumer discrimination tests: Superior performance of constant-saltier-reference DTF and DTFM to same-different tests for discriminating products varying sodium contents. *Food Quality and Preference*. Submitted.

Hough, G., Van Hout, D.H.A., & Kilcast, D. (2006). Workshop Summary: Sensory shelf-life testing, 6th Pangborn Sensory Science Symposium, *Food Quality and Preference*, Volume 17, Issues 7/8, October/December.

Hautus, M.J., Van Hout, D., & Lee, H-S. (2009). Variants of A-Not A and 2-AFC tests: signal detection models. *Food Quality and Preference, 20*, 222-229.

Lee, H-S., & Van Hout, D. (2009). Quantification of sensory and food quality: the R-index analysis. *Journal of Food Science, 74*, 57-64.

Lee, H-S., Van Hout, D.H.A., & Hautus, M. (2007b). Comparison of performance in the A-Not A, 2-AFC, and same-different tests for the flavour discrimination of margarines: the effect of cognitive decision strategies. *Food Quality and Preference, 18*, 920-928.

Lee, H-S., Van Hout, D.H.A., & O'Mahony, M. (2006). Sensory difference tests for margarine: a comparison of R-indices derived from ranking and A-Not A methods considering response bias and cognitive strategies. *Food Quality and Preference, 18*, 675-680.

Lee, H-S., Van Hout, D.H.A., Hautus, M.J., & O'Mahony, M. (2007a). Can the same-different test use a beta criterion as well as a tau criterion? *Food Quality and Preference, 18*, 605-613.

Kim, M-A., Chae, J-E., Van Hout, D.H.A., & Lee, H-S. (2011). Discriminations of the A-Not A difference test improved when “A” was familiarized using a brand image. *Food Quality and Preference, 23*, 3-12.

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

Stocks, M.A., Van Hout, D.H.A., & Hautus, M.J. (2013). Cognitive decision strategies adopted by trained judges when discriminating aqueous solutions differing in the concentration of citric acid. *Journal of Sensory Studies, 28,* 217-229.

Stocks, M.A., Van Hout, D.H.A., and Hautus, M.J. (2014). Cognitive decision strategies adopted by trained judges when discriminating food products differing in the concentration of sucrose and citric acid. *Food Quality and Preference,* submitted.

Van Hout, D., Hautus, M. & Lee, H-S. (2011). Investigation of test performance over repeated sessions using signal detection theory: comparison of three nonattribute-specified difference tests 2-AFCR, A-Not A and 2-AFC. *Journal of Sensory Studies, 26*, 311-321.

Van Hout, D., Dijksterhuis, G.B., & Groenen, P.J.F. (2014). From sensory panels to consumers; utilizing signal detection theory to increase the effectiveness of sensory evaluation. *Journal of the Science of Food and Agriculture*, submitted.

Willems, A.A., Hout, D.H.A., Zijlstra, N. and Zandstra E.H. (2013). Effects of salt labelling and repeated in-home consumption on long-term liking of reduced-salt soups. *Public Health Nutrition* 1-8.

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ERIM Ph.D. Series Research in Management

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

DISSERTATIONS LAST FIVE YEARS

Acciaro, M., *Bundling Strategies in Global Supply Chains*. Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-197-LIS, http://hdl.handle.net/1765/19742

Agatz, N.A.H., *Demand Management in E-Fulfillment*. Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2009-163-LIS, http://hdl.handle.net/1765/15425

Alexiev, A., *Exploratory Innovation: The Role of Organizational and Top Management Team Social Capital*. Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-208-STR, http://hdl.handle.net/1765/20632

Asperen, E. van, *Essays on Port, Container, and Bulk Chemical Logistics Optimization*. **Pr**omoter(s): Prof.dr.ir. R. Dekker, EPS-2009-181-LIS, http://hdl.handle.net/1765/17626

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-Frequency Data*, Promoter(s): Prof.dr.D.J.C. van Dijk, EPS-2013-273-F&A, http://hdl.handle.net/1765/38240

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, http://hdl.handle.net/1765/23670

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2013-278-S&E, http://hdl.handle.net/1765/ 39128

Betancourt, N.E., *Typical Atypicality: Formal and Informal Institutional Conformity, Deviance, and Dynamics*, Promoter(s): Prof.dr. B. Krug, EPS-2012-262-ORG, http://hdl.handle.net/1765/32345

Bezemer, P.J., *Diffusion of Corporate Governance Beliefs: Board Independence and the Emergence of a Shareholder Value Orientation in the Netherlands*. Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-192-STR, http://hdl.handle.net/1765/18458

Binken, J.L.G., *System Markets: Indirect Network Effects in Action, or Inaction,* Promoter(s): Prof.dr. S. Stremersch, EPS-2010-213-MKT, http://hdl.handle.net/1765/21186

Blitz, D.C., *Benchmarking Benchmarks,* Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, http://hdl.handle.net/1765/22624

Borst, W.A.M., *Understanding Crowdsourcing: Effects of Motivation and Rewards on Participation and Performance in Voluntary Online Activities,* Promoter(s): Prof.dr.ir. J.C.M. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-221-LIS, http://hdl.handle.net/1765/ 21914

Budiono, D.P., *The Analysis of Mutual Fund Performance: Evidence from U.S. Equity Mutual Funds,* Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-185-F&A, http://hdl.handle.net/1765/18126

Burger, M.J., *Structure and Cooptition in Urban Networks*, Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.R. Commandeur, EPS-2011-243-ORG, http://hdl.handle.net/1765/26178

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit, Coworker Satisfaction, and Relational Models*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2013-292-ORG, http://hdl.handle.net/1765/1

Camacho, N.M., *Health and Marketing; Essays on Physician and Patient Decision-making*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-237-MKT, http://hdl.handle.net/1765/23604

Carvalho, L., *Knowledge Locations in Cities; Emergence and Development Dynamics*, Promoter(s): Prof.dr. L. van den Berg, EPS-2013-274-S&E, http://hdl.handle.net/1765/ 38449

Carvalho de Mesquita Ferreira, L., *Attention Mosaics: Studies of Organizational Attention*, Promoter(s): Prof.dr. P.M.A.R. Heugens & Prof.dr. J. van Oosterhout, EPS-2010-205-ORG, http://hdl.handle.net/1765/19882

Chen, C.M., *Evaluation and Design of Supply Chain Operations Using DEA*, Promoter(s): Prof.dr. J.A.E.E. van Nunen, EPS-2009-172-LIS, http://hdl.handle.net/1765/16181

Cox, R.H.G.M., *To Own, To Finance, and to Insure; Residential Real Estate Reealed*, Promoter(s): Prof.dr. D. Brounen, EPS-2013-290-F&A, http://hdl.handle.net/1765/1

Defilippi Angeldonis, E.F., *Access Regulation for Naturally Monopolistic Port Terminals: Lessons from Regulated Network Industries,* Promoter(s): Prof.dr. H.E. Haralambides, EPS-2010-204-LIS, http://hdl.handle.net/1765/19881

Deichmann, D., *Idea Management: Perspectives from Leadership, Learning, and Network Theory*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2012-255-ORG, http://hdl.handle.net/1765/ 31174

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2011-232-ORG, http://hdl.handle.net/1765/23268

Diepen, M. van, *Dynamics and Competition in Charitable Giving*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2009-159-MKT, http://hdl.handle.net/1765/14526

Dietvorst, R.C., *Neural Mechanisms Underlying Social Intelligence and Their Relationship with the Performance of Sales Managers,* Promoter(s): Prof.dr. W.J.M.I. Verbeke, EPS-2010-215-MKT, http://hdl.handle.net/1765/21188

Dietz, H.M.S., *Managing (Sales)People towards Performance: HR Strategy, Leadership & Teamwork*, Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2009-168-ORG, http://hdl.handle.net/1765/16081

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-272-LIS, http://hdl.handle.net/1765/38241

Doorn, S. van, *Managing Entrepreneurial Orientation*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-258-STR, http://hdl.handle.net/1765/32166

Douwens-Zonneveld, M.G., *Animal Spirits and Extreme Confidence: No Guts, No Glory*, Promoter(s): Prof.dr. W.F.C. Verschoor, EPS-2012-257-F&A, http://hdl.handle.net/1765/31914

Duca, E., *The Impact of Investor Demand on Security Offerings*, Promoter(s): Prof.dr. A. de Jong, EPS-2011-240-F&A, http://hdl.handle.net/1765/26041

Duursema, H., *Strategic Leadership; Moving Beyond the Leader-follower Dyad*, Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2013-279-ORG, http://hdl.handle.net/1765/ 39129

Eck, N.J. van, *Methodological Advances in Bibliometric Mapping of Science*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-247-LIS, http://hdl.handle.net/1765/26509

Eijk, A.R. van der, *Behind Networks: Knowledge Transfer, Favor Exchange and Performance*, Promoter(s): Prof.dr. S.L. van de Velde & Prof.dr.drs. W.A. Dolfsma, EPS-2009-161-LIS, http://hdl.handle.net/1765/14613

Essen, M. van, *An Institution-Based View of Ownership*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, http://hdl.handle.net/1765/22643

Feng, L., *Motivation, Coordination and Cognition in Cooperatives,* Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2010-220-ORG, http://hdl.handle.net/1765/21680

Gertsen, H.F.M., *Riding a Tiger without Being Eaten: How Companies and Analysts Tame Financial Restatements and Influence Corporate Reputation*, Promoter(s): Prof.dr. C.B.M. van Riel, EPS-2009-171-ORG, http://hdl.handle.net/1765/16098

Gharehgozli, A.H., *Developing New Methods for Efficient Container Stacking Operations*, Promoter(s): Prof.dr.ir. M.B.M. de Koster, EPS-2012-269-LIS, http://hdl.handle.net/1765/ 37779

Gijsbers, G.W., *Agricultural Innovation in Asia: Drivers, Paradigms and Performance,* Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2009-156-ORG, http://hdl.handle.net/1765/14524

Gils, S. van, *Morality in Interactions: On the Display of Moral Behavior by Leaders and Employees*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-270-ORG, http://hdl.handle.net/1765/ 38028

Ginkel-Bieshaar, M.N.G. van, *The Impact of Abstract versus Concrete Product Communications on Consumer Decision-making Processes*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-256-MKT, http://hdl.handle.net/1765/31913

Gkougkousi, X., *Empirical Studies in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. E. Peek, EPS-2012-264-F&A, http://hdl.handle.net/1765/37170

Gong, Y., *Stochastic Modelling and Analysis of Warehouse Operations*, Promoter(s): Prof.dr. M.B.M. de Koster & Prof.dr. S.L. van de Velde, EPS-2009-180-LIS, http://hdl.handle.net/1765/16724

Greeven, M.J., *Innovation in an Uncertain Institutional Environment: Private Software Entrepreneurs in Hangzhou, China*, Promoter(s): Prof.dr. B. Krug, EPS-2009-164-ORG, http://hdl.handle.net/1765/15426

Hakimi, N.A, *Leader Empowering Behaviour: The Leader's Perspective: Understanding the Motivation behind Leader Empowering Behaviour*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2010-184-ORG, http://hdl.handle.net/1765/17701

Hensmans, M., *A Republican Settlement Theory of the Firm: Applied to Retail Banks in England and the Netherlands (1830-2007)*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. S.J. Magala, EPS-2010-193-ORG, http://hdl.handle.net/1765/19494

Hernandez Mireles, C., *Marketing Modeling for New Products,* Promoter(s): Prof.dr. P.H. Franses, EPS-2010-202-MKT, http://hdl.handle.net/1765/19878

Heyden, M.L.M., *Essays on Upper Echelons & Strategic Renewal: A Multilevel Contingency Approach*, Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-259-STR, http://hdl.handle.net/1765/32167

Hoever, I.J., *Diversity and Creativity: In Search of Synergy*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2012-267-ORG, http://hdl.handle.net/1765/37392

Hoogendoorn, B., *Social Entrepreneurship in the Modern Economy: Warm Glow, Cold Feet*, Promoter(s): Prof.dr. H.P.G. Pennings & Prof.dr. A.R. Thurik, EPS-2011-246-STR, http://hdl.handle.net/1765/26447

Hoogervorst, N., *On The Psychology of Displaying Ethical Leadership: A Behavioral Ethics Approach*, Promoter(s): Prof.dr. D. De Cremer & Dr. M. van Dijke, EPS-2011-244-ORG, http://hdl.handle.net/1765/26228

Huang, X., *An Analysis of Occupational Pension Provision: From Evaluation to Redesign,* Promoter(s): Prof.dr. M.J.C.M. Verbeek & Prof.dr. R.J. Mahieu, EPS-2010-196-F&A, http://hdl.handle.net/1765/19674

Hytönen, K.A. *Context Effects in Valuation, Judgment and Choice*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2011-252-MKT, http://hdl.handle.net/1765/30668

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2013-288-LIS, http://hdl.handle.net/1765/ 39933

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics,* Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-222-LIS, http://hdl.handle.net/1765/22156

Jaspers, F.P.H., *Organizing Systemic Innovation*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2009-160-ORG, http://hdl.handle.net/1765/14974

Jiang, T., *Capital Structure Determinants and Governance Structure Variety in Franchising*, Promoter(s): Prof.dr. G. Hendrikse & Prof.dr. A. de Jong, EPS-2009-158-F&A, http://hdl.handle.net/1765/14975

Jiao, T., *Essays in Financial Accounting*, Promoter(s): Prof.dr. G.M.H. Mertens, EPS-2009-176-F&A, http://hdl.handle.net/1765/16097

Kaa, G. van, *Standard Battles for Complex Systems: Empirical Research on the Home Network,* Promoter(s): Prof.dr.ir. J. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-166-ORG, http://hdl.handle.net/1765/16011

Kagie, M., *Advances in Online Shopping Interfaces: Product Catalog Maps and Recommender Systems,* Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2010-195-MKT, http://hdl.handle.net/1765/19532

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promoter(s): Prof.dr. S. Stremersch, EPS-2011-239-MKT, http://hdl.handle.net/1765/23610

Karreman, B., *Financial Services and Emerging Markets,* Promoter(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, http://hdl.handle.net/1765/ 22280

Kwee, Z., *Investigating Three Key Principles of Sustained Strategic Renewal: A Longitudinal Study of Long-Lived Firms,* Promoter(s): Prof.dr.ir. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2009-174-STR, http://hdl.handle.net/1765/16207

Lam, K.Y., *Reliability and Rankings*, Promoter(s): Prof.dr. P.H.B.F. Franses, EPS-2011-230-MKT, http://hdl.handle.net/1765/22977

Lander, M.W., *Profits or Professionalism? On Designing Professional Service Firms*, Promoter(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2012-253-ORG, http://hdl.handle.net/1765/30682

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promoter(s): Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT, http://hdl.handle.net/1765/23504

Larco Martinelli, J.A., *Incorporating Worker-Specific Factors in Operations Management Models,* Promoter(s): Prof.dr.ir. J. Dul & Prof.dr. M.B.M. de Koster, EPS-2010-217-LIS, http://hdl.handle.net/1765/21527

Li, T., *Informedness and Customer-Centric Revenue Management,* Promoter(s): Prof.dr. P.H.M. Vervest & Prof.dr.ir. H.W.G.M. van Heck, EPS-2009-146-LIS, http://hdl.handle.net/1765/14525

Liang, Q., *Governance, CEO Indentity, and Quality Provision of Farmer Cooperatives*m Promoter(s): Prof.dr. G.W.J. Hendrikse, EPS-2013-281-ORG, http://hdl.handle.net/1765/1

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones; New Frontiers in Entrepreneurship Research*, Promoter(s): Prof.dr. A.R. Thurik, Prof.dr. P.J.F. Groenen & Prof.dr. A. Hofman, EPS-2013-287-S&E, http://hdl.handle.net/1765/ 40081

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promoter(s): Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, http://hdl.handle.net/1765/ 22814

Maas, K.E.G., *Corporate Social Performance: From Output Measurement to Impact Measurement,* Promoter(s): Prof.dr. H.R. Commandeur, EPS-2009-182-STR, http://hdl.handle.net/1765/17627

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promoter(s): Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, http://hdl.handle.net/1765/22744

Mees, H., *Changing Fortunes: How China's Boom Caused the Financial Crisis*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2012-266-MKT, http://hdl.handle.net/1765/34930

Meuer, J., *Configurations of Inter-Firm Relations in Management Innovation: A Study in China's Biopharmaceutical Industry*, Promoter(s): Prof.dr. B. Krug, EPS-2011-228-ORG, http://hdl.handle.net/1765/22745

Mihalache, O.R., *Stimulating Firm Innovativeness: Probing the Interrelations between Managerial and Organizational Determinants*, Promoter(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-260-S&E, http://hdl.handle.net/1765/32343

Milea, V., *New Analytics for Financial Decision Support*, Promoter(s): Prof.dr.ir. U. Kaymak, EPS-2013-275-LIS, http://hdl.handle.net/1765/ 38673

Moonen, J.M., *Multi-Agent Systems for Transportation Planning and Coordination*, Promoter(s): Prof.dr. J. van Hillegersberg & Prof.dr. S.L. van de Velde, EPS-2009-177-LIS, http://hdl.handle.net/1765/16208

Nederveen Pieterse, A., *Goal Orientation in Teams: The Role of Diversity*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-162-ORG, http://hdl.handle.net/1765/15240

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in Short-term Planning and in Disruption Management*, Promoter(s): Prof.dr. L.G. Kroon, EPS-2011-224-LIS, http://hdl.handle.net/1765/22444

Niesten, E.M.M.I., *Regulation, Governance and Adaptation: Governance Transformations in the Dutch and French Liberalizing Electricity Industries*, Promoter(s): Prof.dr. A. Jolink & Prof.dr. J.P.M. Groenewegen, EPS-2009-170-ORG, http://hdl.handle.net/1765/16096

Nijdam, M.H., *Leader Firms: The Value of Companies for the Competitiveness of the Rotterdam Seaport Cluster,* Promoter(s): Prof.dr. R.J.M. van Tulder, EPS-2010-216-ORG, http://hdl.handle.net/1765/21405

Noordegraaf-Eelens, L.H.J., *Contested Communication: A Critical Analysis of Central Bank Speech,* Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2010-209-MKT, http://hdl.handle.net/1765/21061

Nuijten, A.L.P., *Deaf Effect for Risk Warnings: A Causal Examination applied to Information Systems Projects*, Promoter(s): Prof.dr. G. van der Pijl & Prof.dr. H. Commandeur & Prof.dr. M. Keil, EPS-2012-263-S&E, http://hdl.handle.net/1765/34928

Nuijten, I., *Servant Leadership: Paradox or Diamond in the Rough? A Multidimensional Measure and Empirical Evidence*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-183-ORG, http://hdl.handle.net/1765/21405

Oosterhout, M., van, *Business Agility and Information Technology in Service Organizations,* Promoter(s): Prof,dr.ir. H.W.G.M. van Heck, EPS-2010-198-LIS, http://hdl.handle.net/1765/19805

Oostrum, J.M., van, *Applying Mathematical Models to Surgical Patient Planning*, Promoter(s): Prof.dr. A.P.M. Wagelmans, EPS-2009-179-LIS, http://hdl.handle.net/1765/16728

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on Bureaucracy and Formal Rules*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, http://hdl.handle.net/1765/23250

Otgaar, A.H.J., *Industrial Tourism: Where the Public Meets the Private,* Promoter(s): Prof.dr. L. van den Berg, EPS-2010-219-ORG, http://hdl.handle.net/1765/21585

Ozdemir, M.N., *Project-level Governance, Monetary Incentives and Performance in Strategic R&D Alliances*, Promoter(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, http://hdl.handle.net/1765/23550

Peers, Y., *Econometric Advances in Diffusion Models*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-251-MKT, http://hdl.handle.net/1765/ 30586

Pince, C., *Advances in Inventory Management: Dynamic Models,* Promoter(s): Prof.dr.ir. R. Dekker, EPS-2010-199-LIS, http://hdl.handle.net/1765/19867

Porras Prado, M., *The Long and Short Side of Real Estate, Real Estate Stocks, and Equity*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2012-254-F&A, http://hdl.handle.net/1765/30848

Potthoff, D., *Railway Crew Rescheduling: Novel Approaches and Extensions,* Promoter(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. L.G. Kroon, EPS-2010-210-LIS, http://hdl.handle.net/1765/21084

Poruthiyil, P.V., *Steering Through: How Organizations Negotiate Permanent Uncertainty and Unresolvable Choices*, Promoter(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. S. Magala, EPS-2011-245-ORG, http://hdl.handle.net/1765/26392

Pourakbar, M. *End-of-Life Inventory Decisions of Service Parts*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2011-249-LIS, http://hdl.handle.net/1765/30584

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promoter(s): Prof.dr. H.R. Commandeur & Prof.dr. H.J.H.M. Claassen, EPS-2013-282-S&E, http://hdl.handle.net/1765/1

Rijsenbilt, J.A., *CEO Narcissism; Measurement and Impact*, Promoter(s): Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, http://hdl.handle.net/1765/ 23554

Roelofsen, E.M., *The Role of Analyst Conference Calls in Capital Markets*, Promoter(s): Prof.dr. G.M.H. Mertens & Prof.dr. L.G. van der Tas RA, EPS-2010-190-F&A, http://hdl.handle.net/1765/18013

Rosmalen, J. van, *Segmentation and Dimension Reduction: Exploratory and Model-Based Approaches*, Promoter(s): Prof.dr. P.J.F. Groenen, EPS-2009-165-MKT, http://hdl.handle.net/1765/15536

Roza, M.W., *The Relationship between Offshoring Strategies and Firm Performance: Impact of Innovation, Absorptive Capacity and Firm Size,* Promoter(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, http://hdl.handle.net/1765/22155

Rubbaniy, G., *Investment Behavior of Institutional Investors*, Promoter(s): Prof.dr. W.F.C. Vershoor, EPS-2013-284-F&A, http://hdl.handle.net/1765/ 40068

Rus, D., *The Dark Side of Leadership: Exploring the Psychology of Leader Self-serving Behavior*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2009-178-ORG, http://hdl.handle.net/1765/16726

Schellekens, G.A.C., *Language Abstraction in Word of Mouth,* Promoter(s): Prof.dr.ir. A. Smidts, EPS-2010-218-MKT, http://hdl.handle.net/1765/21580

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promoter(s): Prof.dr. G.M.H. Mertens, EPS-2013-283-F&A, http://hdl.handle.net/1765/39655

Sotgiu, F., *Not All Promotions are Made Equal: From the Effects of a Price War to Cross-chain Cannibalization,* Promoter(s): Prof.dr. M.G. Dekimpe & Prof.dr.ir. B. Wierenga, EPS-2010-203-MKT, http://hdl.handle.net/1765/19714

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promoter(s): Prof.dr.ir. R. Dekker, EPS-2013-293-LIS, http://hdl.handle.net/1765/1

Srour, F.J., *Dissecting Drayage: An Examination of Structure, Information, and Control in Drayage Operations*, Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-186-LIS, http://hdl.handle.net/1765/18231

Stallen, M., *Social Context Effects on Decision-Making; A Neurobiological Approach*, Promoter(s): Prof.dr.ir. A. Smidts, EPS-2013-285-MKT, http://hdl.handle.net/1765/ 39931

Sweldens, S.T.L.R., *Evaluative Conditioning 2.0: Direct versus Associative Transfer of Affect to Brands,* Promoter(s): Prof.dr. S.M.J. van Osselaer, EPS-2009-167-MKT, http://hdl.handle.net/1765/16012

Tarakci, M., *Behavioral Strategy; Strategic Consensus, Power and Networks*, Promoter(s): Prof.dr. P.J.F. Groenen & Prof.dr. D.L. van Knippenberg, EPS-2013-280-ORG, http://hdl.handle.net/1765/ 39130

Teixeira de Vasconcelos, M., *Agency Costs, Firm Value, and Corporate Investment*, Promoter(s): Prof.dr. P.G.J. Roosenboom, EPS-2012-265-F&A, http://hdl.handle.net/1765/37265

Tempelaar, M.P., *Organizing for Ambidexterity: Studies on the Pursuit of Exploration and Exploitation through Differentiation, Integration, Contextual and Individual Attributes,* Promoter(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-191-STR, http://hdl.handle.net/1765/18457

Tiwari, V., *Transition Process and Performance in IT Outsourcing: Evidence from a Field Study and Laboratory Experiments,* Promoter(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. P.H.M. Vervest, EPS-2010-201-LIS, http://hdl.handle.net/1765/19868

Tröster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://hdl.handle.net/1765/23298

Tsekouras, D., *No Pain No Gain: The Beneficial Role of Consumer Effort in Decision Making*, Promoter(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-268-MKT, http://hdl.handle.net/1765/ 37542

Tzioti, S., *Let Me Give You a Piece of Advice: Empirical Papers about Advice Taking in Marketing,* Promoter(s): Prof.dr. S.M.J. van Osselaer & Prof.dr.ir. B. Wierenga, EPS-2010-211-MKT, hdl.handle.net/1765/21149

Vaccaro, I.G., *Management Innovation: Studies on the Role of Internal Change Agents,* Promoter(s): Prof.dr. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-212-STR, hdl.handle.net/1765/21150

Verheijen, H.J.J., *Vendor-Buyer Coordination in Supply Chains,* Promoter(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2010-194-LIS, http://hdl.handle.net/1765/19594

Venus, M., *Demystifying Visionary Leadership; In Search of the Essence of Effective Vision Communication*, Promoter(s): Prof.dr. D.L. van Knippenberg, EPS-2013-289-ORG, http://hdl.handle.net/1765/ 40079

Verwijmeren, P., *Empirical Essays on Debt, Equity, and Convertible Securities*, Promoter(s): Prof.dr. A. de Jong & Prof.dr. M.J.C.M. Verbeek, EPS-2009-154-F&A, http://hdl.handle.net/1765/14312

Visser, V., *Leader Affect and Leader Effectiveness; How Leader Affective Displays Influence Follower Outcomes*, Promoter(s): Prof.dr. D. van Knippenberg, EPS-2013-286-ORG, http://hdl.handle.net/1765/40076

Vlam, A.J., *Customer First? The Relationship between Advisors and Consumers of Financial Products*, Promoter(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-250-MKT, http://hdl.handle.net/1765/30585

Waard, E.J. de, *Engaging Environmental Turbulence: Organizational Determinants for Repetitive Quick and Adequate Responses*, Promoter(s): Prof.dr. H.W. Volberda & Prof.dr. J. Soeters, EPS-2010-189-STR, http://hdl.handle.net/1765/18012

Wall, R.S., *Netscape: Cities and Global Corporate Networks,* Promoter(s): Prof.dr. G.A. van der Knaap, EPS-2009-169-ORG, http://hdl.handle.net/1765/16013

Waltman, L., *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*, Promoter(s): Prof.dr.ir. R. Dekker & Prof.dr.ir. U. Kaymak, EPS-2011-248-LIS, http://hdl.handle.net/1765/26564

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, http://hdl.handle.net/1765/26066

Wang, Y., *Corporate Reputation Management; Reaching Out to Find Stakeholders*, Promoter(s): Prof.dr. C.B.M. van Riel, EPS-2013-271-ORG, http://hdl.handle.net/1765/ 38675

Weerdt, N.P. van der, *Organizational Flexibility for Hypercompetitive Markets: Empirical Evidence of the Composition and Context Specificity of Dynamic Capabilities and Organization Design Parameters,* Promoter(s): Prof.dr. H.W. Volberda, EPS-2009-173-STR, http://hdl.handle.net/1765/16182

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value of Dutch Firms*, Promoter(s): Prof.dr. A. de Jong, EPS-2013-277-F&A, http://hdl.handle.net/1765/ 39127

Wubben, M.J.J., *Social Functions of Emotions in Social Dilemmas*, Promoter(s): Prof.dr. D. De Cremer & Prof.dr. E. van Dijk, EPS-2009-187-ORG, http://hdl.handle.net/1765/18228

Xu, Y., *Empirical Essays on the Stock Returns, Risk Management, and Liquidity Creation of Banks*, Promoter(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-188-F&A, http://hdl.handle.net/1765/18125

Yang, J., *Towards the Restructuring and Co-ordination Mechanisms for the Architecture of Chinese Transport Logistics*, Promoter(s): Prof.dr. H.E. Harlambides, EPS-2009-157-LIS, http://hdl.handle.net/1765/14527

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promoter(s): Prof.dr. M.B.M. de Koster, EPS-2013-276-LIS, http://hdl.handle.net/1765/1

Zhang, D., *Essays in Executive Compensation*, Promoter(s): Prof.dr. I. Dittmann, EPS-2012-261-F&A, http://hdl.handle.net/1765/32344

Zhang, X., *Scheduling with Time Lags,* Promoter(s): Prof.dr. S.L. van de Velde, EPS-2010-206-LIS, http://hdl.handle.net/1765/19928

Zhou, H., *Knowledge, Entrepreneurship and Performance: Evidence from Country-level and Firm-level Studies,* Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. L.M. Uhlaner, EPS-2010-207-ORG, http://hdl.handle.net/1765/20634

Zwan, P.W. van der, *The Entrepreneurial Process: An International Analysis of Entry and Exit*, Promoter(s): Prof.dr. A.R. Thurik & Prof.dr. P.J.F. Groenen, EPS-2011-234-ORG, http://hdl.handle.net/1765/23422

Erasmus Research Institute of Management - ERiM

MEASURING MEANINGFUL DIFFERENCES

SENSORY TESTING BASED DECISION MAKING IN AN INDUSTRIAL CONTEXT; APPLICATIONS OF SIGNAL DETECTION THEORY AND THURSTONIAN MODELLING

In the 'fast moving consumer goods' industry, results from sensory research form the basis for many important business decisions. Examples of such decisions are whether to launch new products, change existing products, or whether to continue with specific novel technological developments. To make good quality decisions, it is important that the sensory methods used are fast, accurate and deliver robust results.

Signal detection theory and Thurstonian modelling can improve the effectiveness of sensory research, and these theories have been applied to one specific type of methods; sensory difference tests. Sensory difference tests are used to measure small differences between products, and can be used to answer important questions like: *"Are these two products similar in taste?"*, *"Does this new ingredient make the product different?"*, and *"Will our consumers be able to notice the differences?"*

Two signal detection applications have been investigated. The first application is to compare test methods and identify how to optimise them, as there are many methods available that largely differ in performance. With this knowledge, more effective methods can be selected or specifically designed. The second application is to integrate results from different studies to improve the effectiveness of sensory testing in general, for example by relating sensory differences detected by a trained panel *"In Lab"* to differences found by consumers *"In Home"*. Such knowledge can make future studies more predictive of what really matters to consumers, and improve the quality of decision making based on sensory results whilst reducing the amount of testing required.

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERIM PhD Series

Research in Management

Erasmus Research Institute of Management - ERiM
Rotterdam School of Management (RSM)
Erasmus School of Economics (ESE)
Erasmus University Rotterdam (EUR)
P.O. Box 1738, 3000 DR Rotterdam,
The Netherlands

Tel.	+31 10 408 11 82
Fax	+31 10 408 96 40
E-mail	info@erim.eur.nl
Internet	www.erim.eur.nl

Design & layout: B&T Ontwerp en advies (www.b-en-t.nl) Print: Haveka (www.haveka.nl)