ISABEL DE SOUSA AMORIM

NOVEL EFFECT SIZE INTERPRETATION OF MIXED MODELS RESULTS WITH A VIEW TOWARDS SENSORY DATA

LAVRAS - MG

2015

ISABEL DE SOUSA AMORIM

NOVEL EFFECT SIZE INTERPRETATION OF MIXED MODELS RESULTS WITH A VIEW TOWARDS SENSORY DATA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador
Dr. Renato Ribeiro de Lima

LAVRAS - MG
2015

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pela própria autora.

Amorim, Isabel de Sousa.

Novel effect size interpretation of mixed models results with a view towards sensory data / Isabel de Sousa Amorim. – Lavras : UFLA, 2015.

98 p.

Tese – Universidade Federal de Lavras, 2015.

Orientador: Renato Ribeiro de Lima.

Bibliografia.

1. Mixed models. 2. delta-tilde. 3. Sensory analysis. I. Universidade Federal de Lavras. II. Título.

ISABEL DE SOUSA AMORIM

NOVEL EFFECT SIZE INTERPRETATION OF MIXED MODELS RESULTS WITH A VIEW TOWARDS SENSORY DATA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 29 de junho de 2015.

Prof. Dr. Per Bruun Brockhoff	DTU Compute - Statistical Section
Profa. Dra. Ana Carla Marques Pinheiro	DCA - UFLA
Prof. Dr. João Domingos Scalon	DEX - UFLA
Prof. Dr. Júlio Silvio de Sousa Bueno Filho	DEX - UFLA

Prof. Dr. Renato Ribeiro de Lima
Orientador

LAVRAS - MG
2015

A minha mãe e ao meu filho Luís Filipe, com amor.
Ao meu pai, in memorian.
Dedico.

AGRADECIMENTOS

Uma vitória é constituída por muitas etapas e, em cada uma delas, é possível contar com o apoio, incentivo, compreensão, torcida e expectativa de várias pessoas. Uma vitória, antes de tudo, é uma conquista conjunta de professores, orientadores, amigos e familiares. Sou grata a todas as pessoas que, de alguma forma me ajudaram a enfrentar os desafios e alcançar esta vitória.

Agradeço a todos os meus professores que fizeram parte desta jornada. Em especial ao meu orientador, Renato Ribeiro de Lima, pela orientação, pelos ensinamentos e pelo tempo que trabalhamos junto. Agradeço também ao meu orientador estrangeiro, Per Bruun Brockhoff, pela oportunidade de estudar na Dinamarca e por suas contribuições científicas a este trabalho. Obrigada a meus orientadores por acreditarem em meu potencial e me encorajar a enfrentar os desafios.

Agradeço aos colegas e funcionários do Departamento de Exatas, pelo auxílio, ajuda e apoio, em especial à Nádia Ferreira. Agradeço aos amigos de Lavras, em especial à querida amiga Elayne Veiga, pela amizade sincera. Às meninas que dividiram comigo mais que um apartamento, mas também as experiências, frustações e desafios da pós graduação. Nayara Noronha, Elaine Martins, Mírian Souza, Helane França e Priscila Castro, lembrarei sempre da amizade, apoio e companheirismo.

Aos colegas do Departamento DTU Compute, que me fizeram sentir parte deste grupo pesquisa desde que cheguei à DTU, em especial à Alexandra Kuznetsova e Federica Belmonte, agradeço pela amizade. Aos amigos Marco Túlio Alves, Marina Castro e Amanda Lenzi, que compartilharam comigo a experiência de viver na Dinamarca. Juntos superamos a saudade de casa e vivemos momentos inesquecíveis. Meus sinceros agradecimentos aos amigos Vera e Bent Jørgensen que me incentivaram a ir para Dinamarca. Muito obrigada pelo carinho, inspiração e amizade. Agradeço também à família dinamarquesa que me acolheu com tanto carinho. Søren Moesgård e Annegrethe Jørgensen, muito obrigada pelos momentos que tivemos junto e por toda apoio e amizade.

Agradeço com muito carinho Gustavo Araújo, por todo amor e companheirismo. Seu apoio durante essa fase da minha vida foi fundamental. Espero ansiosa pelos novos desafios que enfrentaremos, juntos. A minha mãe, agradeço

pela ajuda, amor, força e apoio para que eu seguisse o meu destino. Aos meus irmãos, Eduardo, João Paulo e Marta e ao meu querido filho Luís Filipe, agradeço pelo amor e compreensão. A querida vovó Belinha, agradeço pelo exemplo de fé e bondade e a querida tia avó Rita, agradeço pelo apoio, carinho e incentivo. Agradeço também a todos os familiares e amigos que me ajudaram, de alguma forma, a concluir esta etapa, em especial ao Tio João Bosco, Tia Celeste, Tia Nonata, Madrinha Elisiária e a querida prima e amiga Cecília Castro. Mesmo simples atitudes, como palavras de incentivo e apoio, foram fundamentais para que eu pudesse chegar aqui.

Finalmente, agradeço a Deus por tudo, principalmente por ter colocado pessoas tão especiais em minha vida...

ACKNOWLEDGEMENTS

An achievement is composed of many steps, and each of them, you can count on the support, encouragement, understanding, twisted and expectations of many people. An achievement, first of all, is a joint achievement of teachers, supervisors, friends and family. I am grateful to all the people who somehow helped me to face the challenges and achieve this conquest.

I would like to thank all my teachers who contributed to my professional and personal growth. In special, I would like to thank my supervisor Renato Ribeiro de Lima, for the support, the teaching during the time we have been working together. Warm and sincere gratitude are also addressed to Per Bruun Brockhoff, my supervisor during the Science without Borders program. Thank you for the opportunity to study in Denmark and for your scientific contributions to this work. Thank you also for believing in my potential and encouraging me to face the challenges.

I thank all the employees and colleagues of the Department of Exact Science, in especial Nádia Ferreira. All my friends in Lavras are especially thanked for the friendship, in especial my dear friend Elayne Veiga. You are very important to me! And my friends who share more than a house, but also experiences, frustrations and challenges of academic life. Nayara Noronha, Elaine Martins, Mírian Souza, Helane França and Priscila Castro, I will always remember your friendship, the inspiration and the pleasant atmosphere where we lived.

I thank all the colleagues and Ph.D students from DTU Compute, who made me feel part of their research group since I came to DTU. Alexandra Kuznetsova and Federica Belmonte are especially thanked for the friendship and inspirations. I also would like to thank you the Brazilians friends Marco Túlio, Marina Castro and Amanda Lenzi, who shared with me a great and exciting experience of living in Denmark. We have had great time together.

Warm gratitude are also addressed to my friends Vera and Bent Jørgensen who encouraged me to go to Denmark. Thank you for always being there for me. And I will always be thankful for the Danish family who welcomed me so warmly in their home. Søren Moesgård and Annegrethe Jørgensen, thank you for the great times we had together, for the inspiring talks and supporting words. You

will always be in my heart.

The warmest gratitude is addressed to Gustavo Araújo, for all your love. Your support during this phase of my life have been so important. I´m looking forward to the new next step, together. Last but not least, I address my warmest thank to my lovely mom. Thank you for your love and encouragement to me to follow my dreams. My brothers, Eduardo, João Paulo and Marta; my dear son Luís Filipe and my grandma Belinha, thanks for your love; my dear aunt Rita, thanks for the encouragement and inspiration. And thanks to all my family and friends who contributed to my personal growth, in especial uncle João Bosco, aunt Celeste, aunt Nonata, godmother Elisiária and my dear cousin and friend Cecília Castro. There are so many people should be acknowledged, I could not mentioned all them here.

Finally, I thank God for everything, but mostly for the amazing people who have crossed my path.

“We know only too well that what we are doing is nothing more than a drop in the ocean. But if the drop were not there, the ocean would be missing something.”

Mother Teresa of Calcutta

ABSTRACT

In sensory studies, the analysis of variance is one of the most often employed statistical methods to study differences between products. However, the analysis of variance often focus just on the p-values. Therefore, it would be valuable to supplement the F-testing with some good measures of overall effect size (ES). In this thesis, a visual tool based on effect size measures is proposed to improve the F-test results interpretations of mixed model ANOVA for sensory data. The basic and straightforward idea is to interpret effects relative to the residual error and to choose the proper effect size measure. The close link between Cohen's d, the effect size in an ANOVA framework, and the Thurstonian (Signal detection) d-prime are used to suggest the delta-tilde barplot as a better visual tool to interpret sensory and consumer data mixed model results. For multi-attribute barplots of F-statistics in balanced settings, this amounts to a simple transformation of the bar heights to get them depicting, what can be seen as approximately the average pairwise d-primes among products levels. The delta-tilde barplot becomes more important for multi-way product models, since the transformation depends on the number of observations within product levels. Then, for extensions into multi-way models, a similar transformation is suggested, in order to make valid the comparison of bar heights for factors with differences in number of levels. The methods are illustrated on a multifactorial sensory profile data set and compared to actual d-prime calculations based on Thurstonian regression modelling through the ordinal R-package. A generic implementation of the method is available on the R-package SensMixed. The use of the delta-tilde barplot can be viewed as good a and relevant additional tools for interpretation of the ANOVA table, particularly in situations with more than a single factor and with several attributes.

Keywords: Mixed Model, delta-tilde, Sensory Analysis.

Guidance Committee: Dr. Renato Ribeiro de Lima - (Supervisor) - UFLA.

RESUMO

A análise de variância é um dos métodos estatísticos mais utilizados para investigar as diferenças entre os produtos em estudos sensoriais. Entretanto, os resultados da análise de variância geralmente focam apenas nos valores-p. Do ponto de vista prático, é relevante complementar os resultados do teste F com alguma estimativa do tamanho do efeito. O objetivo deste trabalho é apresentar um método gráfico baseado nas estimativas de tamanho do efeito (delta-tilde) como uma maneira de aprimorar a interpretação dos resultados do teste F. Para propor o barplot baseado na estimativa delta-tilde, utilizou-se a estreita relação entre o d-prime dos modelos da teoria de detecção de sinais, conhecidos como modelos de Thurstone e a medida d de Cohen, o tamanho do efeito para análise de variância. Para o caso de dados balanceados, a estimativa do delta-tilde é obtida por meio de uma simples transformação da estatística F. A utilização do gráfico baseado no delta-tilde torna-se ainda mais relevante em situações em que os produtos são compostos por mais de um fator, uma vez que a transformação depende do número de observações nos níveis de cada fator. Uma transformação similar é sugerida para modelos multifatoriais, com objetivo de permitir a comparação entre o tamanho do efeito para fatores com níveis diferentes. Apresentou-se um exemplo do uso dos gráficos baseado na estimativa do delta-tilde para interpretar os resultados de uma análise de dados multifatorial. Comparou-se os valores da estimativa do tamanho do efeito (delta-tilde) com o d-prime obtido por meio de uma regressão de Thurstone, utilizando o pacote ordinal, do programa R. Uma implementação geral do método, que permite estimar o tamanho do efeito para casos mais complexos, onde existe desbalanceamento ou dados faltosos, é apresentada no pacote SensMixed, do programa R. Uma das principais vantagens de avaliar os resultados da análise de variância utilizando o gráfico delta-tilde é permitir a comparação do tamanho do efeito entre os fatores, principalmente em situações em que há mais um fator e muitos atributos em estudo.

Palavras-chave: Modelo misto, delta-tilde, Análise sensorial

Comitê Orientador: Dr. Renato Ribeiro de Lima - (Orientador) - UFLA.

List of Figures

1 Bar plot for F values for fixed effects of TVbo data. 48
2 Bar plot based on delta-tilde for fixed effects of TVbo data. 48
3 Bar plot for delta-tilde based on F-statistics from fixed effects model for attribute 7. 50
4 PanelCheck plot: F statistics from 2-way ANOVA for Sound data. 66
5 Barplot for $\sqrt{\chi^2}$ for random effects of Sound data. 74
6 Barplot for delta-tilde estimate of fixed effects of Sound data. . . . 78
7 Barplot for differences of least squares means together with the 95% confidence intervals for Car effect of Att 4. 78
8 Barplot for differences of least squares means together with the 95% confidence intervals for track effect of Att 4. 79
9 Barplot for differences of least squares means together with the 95% confidence intervals for SPL effect of Att 4. 79
10 Barplot for $\sqrt{\chi^2}$ of likelihood ratio test for random-effects for SoundBO data 82
11 Barplot for $\sqrt{F}$ statistics for fixed-effects scaling for SoundBO data 82
12 Barplot for delta-tilde estimates for fixed-effects for SoundBO data 83
13 Barplot for differences of least squares means together with the 95% confidence intervals for Car effect of Att 4 for MAM. 83

List of Tables

1	ANOVA table for the fixed effect model for attribute 7 of TVbo data	49
2	Subset of TVbo data	51
3	ANOVA table for subset of TVbo data	51
4	Categorized data for subset of TVbo data	51
5	Likelihood ratio tests for the random effect and their order of elimination for the automated analysis of SoundBO data.	70
6	F-tests for the fixed-effects and their order of elimination of the automated analysis for SoundBO data.	70
7	$\sqrt{\chi^2}$-statistics for LRT for random-effects with significance levels for the SoundBO data.	73
8	F-test for the fixed effects for SoundBO data	75

CONTENTS

1	**INTRODUCTION**	16
1.1	**Challenges for sensory and consumer data analysis**	16
1.2	**Using effect size to improve data analysis**	16
1.3	**Overview of the thesis**	17
2	**STATISTICS FOR SENSORY AND CONSUMER DATA**	18
2.1	**Sensory Science and Sensometrics**	18
2.2	**Analysis of Variance Model for Sensory Profiling**	20
2.2.1	**Mixed Analysis of Variance Model**	20
2.2.2	**Mixed Assessor Model**	23
2.2.3	**Extending the Mixed Assessor Model**	25
2.3	**P-values and Effect Size**	27
2.4	**Thurstonian d' and Cohen's d**	30
3	**DELTA-TILDE INTERPRETATION OF STANDARD LINEAR MIXED MODELS RESULTS**	32
3.1	**Introduction**	33
3.2	**Cohen's d and d-prime - important effect size measures**	35
3.3	**Methods**	39
3.4	**The sample estimation of the $\tilde{\delta}$ ES measures**	41
3.4.1	**The independent two- and multi-group one-way ANOVA case**	41
3.4.2	**Bias of sample estimates and possible bias corrections**	42
3.4.3	**Some standard mixed model sensory and consumer cases**	43
3.4.4	**Back transforming F-statistics more generally**	44
3.4.5	**More general mixed models**	45
3.5	**Examples**	45
3.5.1	**Example 1: Multi-way product structures in sensory profile data**	46
3.5.2	**Example 2: Comparison with d-prime from Thurstonian model - simple example**	50

3.6 Discussion 52
4 TOOLS FOR MIXED MODELLING OF SENSORY DATA . . 53
4.1 Introduction 53
4.2 Overview of the recently developed tools for fitting mixed models to sensory data 54
4.2.1 PanelCheck software 54
4.2.2 lmerTest package 56
4.2.3 SensMixed package 60
4.3 Example: mixed model analysis of sensory study 62
4.3.1 Sensory study of car audio system 63
4.3.2 One-way product analysis using PanelCheck 63
4.3.3 3-way product analysis using lmerTest 65
4.3.4 3-way product analysis using SensMixed 69
4.3.5 Advanced mixed modelling for sensory data using SensMixed 77
4.4 Discussion 84
5 CONCLUSIONS AND FUTURE PERSPECTIVES 86
REFERENCES 88
Appendix A 94
Appendix B 96

1 INTRODUCTION

1.1 Challenges for sensory and consumer data analysis

Data analysis within the sensory and consumer science fields can be particularly challenging due to use of human as the measurement instrument. Understanding how responses change due to product differences versus change due to subject differences is important. Analysis of variance (ANOVA) is one of the most often employed statistical tools to study differences between products when they are scored by either categorical rating (ordinal) scales and/or unstructured line scales (NÆS; BROCKHOFF; TOMIC, 2010). A number of relevant *post hoc* analysis, also called multiple comparison tests, usually characterizes analysis of variance based data analysis within the sensory field. However, it is still valuable to be able to supplement the initial overall ANOVA F-testing, often with highest focus on the p-values with some good measures of overall effect size.

1.2 Using effect size to improve data analysis

Effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect. It can be as simple as a mean, a percentage increase, or a correlation; or it may be a standardized measure of a difference, a regression weight, or the percentage of variance accounted for. For a two-group setting, the ES quantifies the size of the difference between two groups, and may therefore be said to be a true measure of the significance of the difference (COE, 2002).

An important class of ES measures is defined by using the standardized effect size. In this class is included the Cohen's d, which is the difference measured in units of some relevant standard deviation (SD) (CUMMING; FINCH, 2005). The main purpose of the present study is to improve research interpretation of the results of standard sensory and consumer data mixed model ANOVA suggesting a visual tool based on effect size measures.

1.3 Overview of the thesis

The remainder of this thesis is organized as follows. In Chapter 2 a literature review is presented. In chapter 3 we define the effect size delta-tilde and discuss why effects size are good measure to improve ANOVA interpretation. Plots based on effect size delta-tilde are suggested to supplement the initial overall ANOVA F-testing. The method proposed here is illustrated on a multifactorial sensory profile data set and the delta-tilde proposed here is compared with the actual d-prime based on Thurstonian modelling. Chapter 4 is a review about tools recently developed to improve sensory and consumer data analysis. We focus on the new open source softwares as PanelCheck (NOFIMA; ÅS, 2008), ConsumerCheck (TOMIC et al., 2015) and the two new R-packages lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a) and SensMixed (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014b). These packages provide nice and visual multi-attribute plots with the purpose of improve the interpretation of the results of the (mixed) ANOVA in a best possible way. A plot based on delta-tilde effect size is one of the options provided on SensMixed R-package. In the general discussion of Chapter 5, we will discuss the main contributions and provide recommendations for further researches.

2 STATISTICS FOR SENSORY AND CONSUMER DATA

2.1 Sensory Science and Sensometrics

Sensory science is a cross-disciplinary scientific field dealing with human perception of stimuli and the way they act upon sensory input. In this field the use of humans as measurement instrument play an important role in product development and user-driven innovation in many industries (BROCKHOFF, 2011).

The term sensory science is frequently used to comprise all types of test where human senses are used. According with Næs, Brockhoff and Tomic (2010), the difference between sensory tests using a trained panel and tests using consumers is the way they are used. Sensory panel studies is either used for describing the degree of product similarities and differences in terms of a set of sensory attributes, so-called sensory profiling, or for detecting differences between products, so-called sensory difference testing (MEILGAARD; CIVILLE; CARR, 2006; O'MAHONY, 1986). For consumer studies, the products are tested by a representative group of consumers who are asked to assess their degree of liking, their preferences or their purchase intent for a number of products. These tests are often called hedonic or affective tests (LAWLESS; HEYMANN, 2010).

Sensory and consumer data are produced and applied as the base for decision making in food industry and within many no-food areas, for instance in the automotive, fragrance, mobiles phones, high end TV and audio industries or whatever. The production and interpretation of such data is also an integral part of production development and quality control. The development and application of statistics and data analysis within this area is called sensometrics (BROCKHOFF, 2011), a scientific field that grew out of and is still closely linked, to sensory science.

Sensometrics began with inferential statistics during the 1930s and continued in that direction into 1940s and 1950s. At the early development of the field, one of the key functions of sensory analysis was to test products and to provide an informed opinion about whether or not the product met or failed the acceptance standards (MOSKOWITZ; SILCHER, 2006). We can say that Sensometrics began its

life as simple tests of significance and treatment effects, been strongly influenced on the one hand by trends in experimental psychology and on the other by agricultural statistics. Indeed, Fisher′s classic paper on a "The Mathematics of a Lady Tasting Tea" (FISHER, 1956) illustrates this type of interrelationship of statistics (or at least quantitative methods) and sensory analysis.

Rose Marie Pangborn (1932 - 1990) is considered one of the pioneers of sensory of food and there is an international scientific conference in sensory science, the Pangborn Sensory Science Symposium, dedicated to her memory. She is co-author of the first exposition of the modern sensory science (AMERINE; PANGBORN; RESSSLER, 1965) which served as the definitive textbook for an entire generation of sensory scientists.

The beginning workers in the food industry were occasionally in contact with psychologists who studies the senses and had developed techniques for assessing sensory functions (MOSKOWITZ, 1983). While psychologists focus on understanding how the human sense works, the sensory scientists focus on better understanding of how the senses react during food intake, and also how human senses can be used in quality control and innovative product development, both using statistical methods. In that way, Sensometrics - the "metric" side of sensory science field (BROCKHOFF, 2011), received the influence of different fields mainly statistics, chemometrics and experimental psychology. The use of Thurstonian modelling to form the theoretical basis for sensory discrimination protocols and models for preferential choice (BROCKHOFF; CHRISTENSEN, 2010) are good examples of what can be realized when there is interaction between experimental psychology, statistics and food science methods. According with Lawless and Heymann (2010), differences in language, goals, and experimental focus are some of the difficulties that explain why this interchanges were not more sustained and productive.

We can be optimistic about the future of the sensory science. Judging the number of industries and the many companies that are expanding their sensory test capabilities the future of sensory science is bright. According with Stone and Sidel (2004), there is no question that sensory evaluation is a profitable investment. We could say it in terms of one-to-ten ratio; that is, for every dollar invested in

sensory, it will return ten to the business. The development and use of predictive models of consumer-product behaviour has had and will continue to have a salutary impact, not only because of this specific and immediate value to a company but also because of effectiveness in demonstrating a higher degree of sophistication than was realized possible for sensory evaluation.

2.2 Analysis of Variance Model for Sensory Profiling

Sensory profiling or so-called descriptive sensory analysis is the most sophisticated of the methodologies available to the sensory professional (LAWLESS; HEYMANN, 2010; STONE; SIDEL, 2004) and probably the most important method in sensory analysis (NÆS; BROCKHOFF; TOMIC, 2010). These techniques are used for describing products in terms of the perceived sensory attributes and identify differences between products by the use of trained sensory assessors. This is the case for both product development situations and for quality control.

In sensory profiling, a group of trained assessors, so-called sensory panel, develop a test vocabulary (defining attributes) for the product category and rate the intensity of these attributes for a set of different samples within the category. Thus, a sensory profile of each product is provide for each of the assessors, and most often this is replicated (LAWLESS; HEYMANN, 2010). According with Lawless and Heymann (2010) descriptive analysis techniques should never be used with consumers, because in all descriptive methods, the assessor should be trained at the very least to be consistent and reproducible.

2.2.1 Mixed Analysis of Variance Model

The Analysis of Variance (ANOVA) is the most appropriate statistical procedure for analysing data from descriptive sensory tests and other sensory tests where more than two products are compared using scaled responses (LAWLESS; HEYMANN, 2010; STONE; SIDEL, 2004). In this situation, the proper analysis of variance will typically evaluate the statistical significance of product differences by using the assessor-by-product interaction as error structure (LAWLESS; HEY-

MANN, 2010). In statistics, this is generally called a mixed effect model, as both fixed-effect (products) and random-effects (assessor and assessor-by-product interaction) are presented in the modelling and analysis approach (KUZNETSOVA et al., 2015).

Let Y_{ijk} be the k^{th} replicate evaluation of product j, $j = 1, 2, ..., J$, by assessor i, $i = 1, 2, ..., I$. The simple 2-way mixed analysis of variance model is given by

$$Y_{ijk} = \mu + a_i + \nu_j + g_{ij} + \varepsilon_{ijk}, \tag{1}$$

where a_i is the assessor effect, the ν_j is the product effect the g_{ij} is the assessor-by-product interaction and ε_{ijk} is the residual error. The assessor effect and therefore also the assessor-by-product interaction are assumed random effects. That means

$$\begin{gathered} a_i \sim N(0, \sigma^2_{assessor}), \\ g_{ij} \sim N(0, \sigma^2_{assessor \text{x} product}), \\ \varepsilon_{ijk} \sim N(0, \sigma^2_{error}), \end{gathered}$$

where all the random-effects are independent of each other.

The model (1) is frequently used as the basis for doing univariate statistical analysis of sensory profile data (NÆS; BROCKHOFF; TOMIC, 2010). The hypothesis of interest in model (1) is

$$H_0 : \nu_1 = \nu_2 = ... = \nu_j = 0 \tag{2}$$

which corresponds to no average differences between the main effect for the products. The alternative hypothesis is usually that, at least, two means of the products are different. In the mixed model (1), the null hypothesis (2) refers to the average differences between products for the whole population of potential assessors rather than to specific assessors. If the H_0 is rejected, one should be interested in which products are responsible for the rejection.

ANOVA includes a particular form of null hypothesis statistical testing (NHST) used to identify and to quantify the factors that are responsible for the variability of the response. The null hypothesis for ANOVA is that the means of the factors are the same for all groups. The alternative hypothesis is that at least one mean is different from the others. An F-statistic is obtained in the ANOVA

and the F distribution is used to calculate the p-value. A predefined α, called significance level, is considered as decision criterion, and the null hypothesis is rejected if the p-value is smaller than the value of the α.

In other words, NHST assesses the probability of obtaining the sample data (D) if the null hypothesis (H_0) is true, that is, $p(D|H_0)$. If the $p(D|H_0)$ is sufficiently small (smaller than the decision criterion, α), the null hypothesis will be considered not viable and will be rejected. The rejection of the null hypothesis indicates that the random sampling variability is the unlikely explanation for the observed statistics. We could simply interpret the rejection of the null hypothesis as: given the observed magnitude of difference between the two samples it is highly unlikely that sampling error could have been the cause for the observed data (FAN, 2010).

Considering the assessors as random effects is a proper approach in sensory field (LAWLESS; HEYMANN, 2010) where the main interest is in the population of assessors rather than the actual assessors at hand. According to (KUZNETSOVA et al., 2015) this means that we want to know the variation among assessors rather than estimates of effects of each assessors and to be able to properly account for that. It statistically means that in model (1) we are interested in estimating $\sigma^2_{assessor}$ and $\sigma^2_{assessor\text{X}product}$.

Still an ongoing discussion whether sensory assessors are ever considered as fixed effects (LAWLESS, 1998; LUNDAHL; MACDANIEL, 1988; NÆS; LANGSRUD, 1996; O'MAHONY, 1986). Even though behavioural science suggests that human beings are random effects as they are used in sensory science, unfortunately, the fixed effect model persisted in the literature (LAWLESS; HEYMANN, 2010). According with Brockhoff, Schlich and Skovgaard (2015) both types of analysis can be done with the proper interpretations of the results. In fact, Næs and Langsrud (1996) showed that in the case with no interaction, the two approaches give the same results. However, in the situations with significant interactions between assessors and products, considering assessors as fixed effects may lead to a conclusion that differences between products are larger than they really are. For that reason, the most appropriate assumption is to treat assessors as a random effect. Furthermore, as pointed out in Brockhoff, Schlich and Skovgaard (2015), the ran-

dom assessor type of interpretation resembles better the usual purpose of performing a description sensory experiment: to achieve at some results for the products in question that may be generalized to a larger setting than merely the assessors that are in panel.

By considering assessors as random effect, we have the appropriate F-test for investigating product difference recommended by Lawless and Heymann (2010):

$$F_{Prod} = \frac{MS_{(Product)}}{MS_{(Assessor \times Product)}}$$

According with Næs, Brockhoff and Tomic (2010) if an attribute has no significant main product effect or interaction, it can be safely claimed that the panel as a whole is not able to distinguish between the products for this attribute.

2.2.2 Mixed Assessor Model

Lawless and Heymann (2010) pointed out that the use of an interaction term as the denominator for error has important consequences for the analysis and its sensitivity. It was also indicated by Næs and Langsrud (1996) that a large portion of the interaction effect may be due to individual differences in use of the scale. In the standard mixed model analysis of variance (1) approach, this difference has entered the resulting assessor-by-product error term but pooled together with potential disagreement variability. Another approach is to use the Mixed Assessor Model (MAM) recently developed by Brockhoff, Schlich and Skovgaard (2015) which is based on doing a more elaborate modelling of the interaction. In the MAM, the interaction term is modelling the potential individual differences between the assessors in their scoring of the product differences. This includes as well differences in individual ranges of scale use (scale effect), as the real differences in perception of product differences (disagreement effect) (BROCKHOFF; SCHLICH; SKOVGAARD, 2015). If strong scaling effect are present, the assumptions behind the general ANOVA approach, where the interaction are assumed to be independently distributed, become less realistic. Therefore, MAM gives a more powerful analysis by removing the scaling difference from the interaction term.

In the MAM, the scaling part of the interaction is modeled by incorporating the product effect also as a covariate. In this case, the covariate identify and remove the scaling heterogeneity from the interaction term, and it should not be seen as an analysis of covariance in the usual meaning (BROCKHOFF; SCHLICH; SKOVGAARD, 2015). The Mixed Assessor Model (MAM) is given by

$$Y_{ijk} = \mu + a_i + \nu_j + \beta_i x_j + d_{ij} + \varepsilon_{ijk} \tag{3}$$

with

$$\begin{gathered} a_i \sim N(0, \sigma^2_{assessor}), \\ d_{ij} \sim N(0,\sigma^2_D), \\ \varepsilon_{ijk} \sim N(0,\sigma^2), \end{gathered}$$

where a_i is the assessor main effect, $i = 1, 2, ..., I$, the ν_j the product main effect, $j = 1, 2, ..., J$, $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ are the centered product averages inserted as a covariate, and hence β_i is the individual (scaling) slope (with $\sum_{i=1}^{I} \beta_i = 0$), the d_{ij} takes the role of the random interaction term g_{ij} in the standard mixed model in (1) but now the term captures interactions that are not scale differences hence "disagreements". Brockhoff, Schlich and Skovgaard (2015) has shown that MAM produces valid and improved hypothesis tests for as well overall product differences as *post hoc* product difference testing restricted to the two-way setting.

The main overall hypothesis of interest in model (3) is that there is no product difference:

$$H_0 : \nu_1 = \nu_2 = ... = \nu_j = 0 \tag{4}$$

The advantage of using the MAM rather than the simple mixed model (1) is that it is taken into account that the interaction can be due to scaling difference or due to disagreements. The MAM removes the scaling effect from the interaction term. The consequence for the test of product differences is that the disagreement mean square becomes the one to use in the denominator, improving hypothesis tests for product effects. The appropriate F-statistic for investigating product differences became (BROCKHOFF; SCHLICH; SKOVGAARD, 2015):

$$F_{Prod} = \frac{MS_{(Product)}}{MS_{(Disagreement)}}$$

By removing the scaling effect from the interaction term, the scaling part of the variance structure in the model disappear under the null hypothesis. This is exactly why this novel approach proposed by Brockhoff, Schlich and Skovgaard (2015) provides increased power for detecting product differences: the error used for deciding about product differences has been cleaned out for potential scaling structure.

In addition, Brockhoff, Schlich and Skovgaard (2015) present the F-tests for scaling differences and disagreement:

$$F_{Scaling} = \frac{MS_{(Scaling)}}{MS_{(Disagreement)}}$$

$$F_{disagreement} = \frac{MS_{(Disagreement)}}{MS_{(Error)}}$$

The former investigates whether the scaling are different from individual to individual. And the hypothesis to test the presence of scaling is given by:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_j = 1$$

2.2.3 Extending the Mixed Assessor Model

Brockhoff, Schlich and Skovgaard (2015) showed that MAM produces valid and improved hypothesis tests for as well overall product differences as *post hoc* product difference testing. However, sensory studies are frequently made in replicates/sessions and the MAM given by Equation (3) considers a rather simple 2-way structure. Therefore, it is sensible to consider more complex structures such as 3-way, where the replicate/session effect is also accounted for. Kuznetsova et al. (2015) presented an extended version of MAM, where scaling effect can be part of a more complicated linear mixed effects model and provide a tool to construct

and visualize the results. The 3-way linear mixed assessor model presented by Kuznetsova et al. (2015) is specified in the following form:

$$y_{ijkl} = \mu + a_i + \nu_j + \beta_i x_j + d_{ij} + r_k + ar_{ik} + \nu r_{jk} + \varepsilon_{ijkl} \tag{5}$$

$$\begin{gathered} a_i \sim N(0, \sigma^2_{assessor}), \\ d_{ij} \sim N(0, \sigma^2_{disagreement}), \\ r_k \sim N(0, \sigma^2_{replicate}), \\ ar_{ik} \sim N(0, \sigma^2_{assessor \times replicate}), \\ \nu r_{jk} \sim N(0, \sigma^2_{product \times replicate}), \\ \varepsilon_{ijk} \sim N(0, \sigma^2) \end{gathered}$$

From the Equation (5) we may notice that three more random effects are included in the MAM: r_k corresponding to the replication/session and the remains effects ar_{ik} and νr_{jk} corresponding to the interactions between assessor and replication and assessor and product, respectively. The product effect is represented again by ν_j, the assessor main effect is represented by a_i and $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ are the centered product averages inserted as a covariate, and β_i is the individual (scaling) slope (with $\sum_{i=1}^{I} \beta_i = 0$) and the d_{ij} is the disagreement term. Kuznetsova et al. (2015) also proposed an extended versions of MAM where a possible multi-way product structure can be accounted for together with the three-way error structure, where a replicate effect is also accounted for.

The main overall hypothesis of interest in model (5) is that no product difference:

$$H_0 : \nu_1 = \nu_2 = ... = \nu_j = 0, \tag{6}$$

and the appropriate F-statistic for investigating this hypothesis is given by:

$$F_{Prod} = \frac{MS_{(Product)}}{MS_{(Disagreement)}}$$

The ANOVA F-table, where product effects are found significant, is complemented with *post hoc* test, also called multiple comparison tests. We will show in the next section that it is still valuable to be able to supplement the initial over-

all ANOVA F-testing, often with highest focus on the p-values with some good measures of overall effect size.

2.3 P-values and Effect Size

Analysis of Variance (ANOVA) provides a very sensitive tool for seeing whether treatment variable such as changes in ingredients, processes, or packaging had an effect on the sensory properties of the product (LAWLESS; HEYMANN, 2010). The Analysis of Variance procedure includes a particular form of null hypothesis statistical testing (NHST) used to identify and quantify the factors that are responsible for the variability of the response. The null hypothesis for ANOVA is that the means of the factors are the same for all groups. To test this null hypothesis, the variance or squared deviations due to each factor and the variance or squared deviation due to error are estimated. The error can be thought of as the variability caused by the variables that are not under control in an experiment. Then a ratio of the factor variance and the error variance is constructed. This ratio follows the distribution of an F-statistics, which is used to obtain the p-value. A predefined α, called significance level, is considered as decision criterion, and the null hypothesis is rejected if the p-value is smaller than the value of the α. A significant F-ratio for a given factor implies that, at least, one mean is different from the others.

The NHST is a direct form and an easy way to conclude about the statistical significance of a factor, by considering a significance level and a p-value. However, no single statistical concept is probably more often misunderstood and so often abused as the obtained p-value (HUBBARD; LINDSAY, 2008). Most importantly, the p-value does not provide us a crucial piece of information: the magnitude of an effect of interest. According with Lawless and Heymann (2010) it is important to keep in mind that the p-value is based on a hypothetical curve for the test statistic that is obtained under the assumption that the null hypothesis is true. Therefore, the obtained p-value is taken from the very situation that we are trying to reject or eliminate as a possibility (LAWLESS; HEYMANN, 2010). Therefore, the NHST only informs us of the probability of the observed or more extreme data given the null hypothesis true.

As pointed out by Lawless and Heymann (2010) the NHST by itself is somewhat impoverished manner of performing scientific research. It can be thought of as a starting point or a kind of necessary hurdle that is a part of experimentation in order to help rule out the effect of chance. However, it is not the end of the story, only the beginning. In addition to statistical significance, the sensory science must always describe the effect. The recommendation reporting results (COHEN, 1990, 1992, 1994; DEVANEY, 2001; FAN, 2010; GRISSOM; KIM, 2012; KELLEY; PREACHER, 2012; SUN; PAN; WANG, 2010) is that instead of only reporting p-values, the researchers should also provide estimates of effect size. According with Cohen (1990) the purpose of the research should be to measure the magnitude of an effect rather than simply its statistical significance; thus, reporting and interpreting the effect size is crucial. Therefore, researchers should consider both p-value and effect size (COHEN, 1990).

The term "effect size" (ES) is a name given to a family of indices that measure the magnitude of a treatment. It can be simple as a mean, a mean difference, or a correlation; or it may be a standardized measure of a difference, a regression weight, or the percentage of variance account for. These indices sometimes are called effect size measurement or effect statistics. An effect size can also refers to the actual values calculated from certain effect size measurement, i.e., the effect size value. In other words, an effect size is a relevant interpretation of an estimated magnitude of treatment effect obtained from an effect size statistics. This is sometimes referred as the practical importance of the effect.

An important class of ES measures is defined by using the standardized effect size. In this class is included the Cohen's d, which is the difference measured in units of some relevant standard deviation (SD) (CUMMING; FINCH, 2005). Cohen's d is the ES index for the t-test of the difference between independent means expressed in units of (i.e., divided by) the within-population standard deviation, which is given by:

$$d = \frac{\mu_a - \mu_b}{\sigma}$$

where μ_a and μ_b are independent means and σ is the within-population standard deviation.

In other words, the effect size for the general t-test can be seen as the distance between the mean of the t-distribution (usually zero) under the null hypothesis and the means of the t-distribution under some fixed alternative hypothesis. In sensory discrimination test, we can think of this as the distance between the means of a control product and the mean of a test product under the alternative hypothesis, in standard deviation units. Then the values of the effect size can be set on the basis of d' (d-prime) estimates from signal detection theory (LAWLESS; HEYMANN, 2010), represented by δ when it refers to the population effect size (ENNIS, 1993). As standardized metric, any effect reported in form of d' can be compared with any other.

There are several effect size measures to use in the context of an F-test for ANOVA. Cohen (1992) defined the effect size for one-way ANOVA as the standard deviation of the K population means divided by the common within-population standard deviation:

$$f = \frac{\sigma_m}{\sigma} \tag{7}$$

where σ_m is the standard deviation of the K population mean and σ is the within-population standard deviation.

In fact, the analysis of variance is a way to address multiple treatments or levels, i.e., to compare several means at the same time, while in the t-test we compare only two means. That means, there is an obvious relationship between F and t-statistics: in a simply two-level experiment with only one variable, the F-statistics is simply the square of the t-value (LAWLESS; HEYMANN, 2010). As t-statistics, the F ratio can be seen as an effect size itself. However, the F-statistic is not the best measure of effect size as it depends on the number of observations for each product. In addition, the various ANOVA mixed models, that we often use for such analysis also complicates the relative effect size handling. For example in mixed models, different effects may have different noise structures, that is, different factors may be tested using different F-test denominators. The main aim of this thesis is to present a visual tool to interpret mixed models ANOVA results based on a more proper effect size measure.

A very similar measure of standard ES for ANOVA is the root-mean-

square standardized effect (Ψ) presented by Steiger (2004). Considering the one-way, fixed-effects ANOVA, in which K means are compared for equality, and there are n observations per group the root-mean-square standardized effect is defined by

$$\Psi = \sqrt{\frac{\frac{1}{K-1}\sum_{i=1}^{K}(\mu_i - \mu)^2}{\sigma^2}} \tag{8}$$

where σ^2 is the mean square error. In fact, this could be just an interpretation of what the Cohen's f really is using $K-1$ for expressing the standard deviation as opposed to using K as others might do.

2.4 Thurstonian d' and Cohen's d

Cohen's d, the effect size used to indicate the standardized difference between two means, has a close link with the Thurstonian d', a signal-to-noise ratio from Signal Detection Theory (SDT), which is widely used in sensory science. The Thurstonian d' is the statistic used to estimate the value for δ, the fundamental measure of sensory difference in the Thurstonian model. Mathematically speaking Cohen's d and the Thurstonian d' are exactly the same: the difference between two means relative to a standard deviation. Only the contexts are usually different. The Thurstonian d' is a key parameter quantifying the sensory difference between the stimuli for sensory discrimination test (e.g. the duo-trio, triangle, 2-AFC, 3-AFC) (MEILGAARD; CIVILLE; CARR, 2006), while the Cohen's d is used as a simply way to quantifying the size of the difference between two groups (COE, 2002).

Psychophysics, a branch of experimental psychology devoted to studying the relationships between sensory stimuli and human responses, has a strong influence in sensory science. Perhaps the most widely applied and influential theory in all of the experimental psychology has been signal detection theory (SDT) (LAWLESS; HEYMANN, 2010). SDT approach was first introduced in sensory science with focus on exploring the factors influencing the perceptual process that integrates the information from the senses and the decision process (O'MAHONY, 1972, 1979), leading to more effect test designs (O'MAHONY, 1995). After this

the focus shifted towards understanding and optimizing the decision processes in sensory tests leading to the development of more effective tests that are more predictive of consumer's reality (HAUTUS; O'MAHONY; LEE, 2008).

The framework Thurstonian modelling (THURSTONE, 1927) is a more elaborated model of human behaviour used to understand better the results observed in discrimination test. The Thurstonian approach is responsible for the biggest impact on the development of sensory difference discrimination methods. The statistical methods needed for analysing such data can be found among methods based on the binomial distribution and standard methods for analysing tables of counts, since this kind of tests produce binary data. As pointed out by Ennis (1993) one of the weaknesses of working on the count scale is that it is a test protocol dependent: the number of expected correct answers for the same products depend heavily on which test that is carried out. By transforming the number of correct answers into an estimate of the underlying (relative) sensory difference, the Thurstonian model estimates the size of a sensory difference from a particular test, the so-called d-prime (d'). This measure can be seen as generalized measure of sensory difference that expresses size of sensory differences. Since the d' is independent of the test method used, it 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 (HOUT, 2014).

Although the Thurstonian approach is the most well-known for its use for sensory discrimination test protocols with binary or ordered categorical outcomes, it has also been suggested and used for ratings data, see e.g. Ennis (1999) and Warnock, Shumaker and Delwiche (2006) and also in the context of multivariate analysis of ratings data as e.g. probabilistic multidimensional scaling, cf. MacKay and Zinnes (1986). Brockhoff and Christensen (2010) and Christensen, Cleaver and Brockhoff (2011) showed how the Thurstonian approach in many cases can be viewed as and embedded into the so-called generalized linear model and/or ordinal regression theory and framework. One benefit of this is the ability to handle regression and ANOVA type analysis within the framework of a Thurstonian approach, where otherwise the most common Thurstonian approach would be to do repeated one- and two-sample computations on various subsets of the data.

3 DELTA-TILDE INTERPRETATION OF STANDARD LINEAR MIXED MODELS RESULTS

Isabel de Sousa Amorim - Universidade Federal de Lavras [3]
Alexandra Kuznetsova - Technical University of Denmark
Per Bruun Brockhoff - Technical University of Denmark
Søren Bech - Bang & Olufsen A/S, Struer and Aalborg University, Denmark
Renato Ribeiro de Lima - Universidade Federal de Lavras

Abstract

We utilize the close link between Cohen's d, the effect size in an ANOVA framework, and the Thurstonian (Signal detection) d-prime to suggest better visualizations and interpretations of standard sensory and consumer data mixed model ANOVA results. The basic and straightforward idea is to interpret effects relative to the residual error and to choose the proper effect size measure. For multi-attribute bar plots of F-statistics this amounts, in balanced settings, to a simple transformation of the bar heights to get them transformed into depicting what can be seen as approximately the average pairwise d-primes between products. For extensions of such multi-attribute bar plots into more complex models, a similar transformation is suggested and becomes more important as the transformation depends on the number of observations within factor levels, and hence makes bar heights better comparable for factors with differences in number of levels. For mixed models, where in general the relevant error terms for the fixed effects are not the pure residual error, it is suggested to base the d-prime-like interpretation on the residual error. The methods are illustrated on a multi-factorial sensory profile data set and compared to actual d-prime calculations based on Thurstonian regression modelling through the ordinal package. For more challenging cases we offer a generic implementation of the method as part of the R-package SensMixed.

Keyword: Effect Size, Analysis of Variance, F test, d-prime

[3] Paper submitted to Food Quality and Preference

3.1 Introduction

Data analysis within the sensory and consumer science fields can be particularly challenging due to use of humans as the measurement instrument. Understanding how responses change due to product differences versus change due to subject differences is important. Analysis of variance (ANOVA) is one of the most often employed statistical tools to study differences between products when they are scored by either categorical rating (ordinal) scales and/or unstructured line scales. If for instance one finds that the main product effect is significant, one will be interested in knowing more about which products that are different from each other. To complement the ANOVA F-table, *post hoc* tests are performed. These procedures, also called multiple comparison tests, are generally based on adjusting the critical values of the individual tests in such a way that the overall significance level is controlled. An often used of these tests is the Tukey's test based on comparing differences between pair of means with an adjusted critical value. Other methods that can be used for *post hoc* analysis are the Bonferroni method, Newman-Keul's test and Ducan's test (NÆS; BROCKHOFF; TOMIC, 2010).

Data analysis based on analysis of variance within the sensory field is usually characterized by a number of such relevant *post hoc* analysis. To some extend this then handles the effect interpretation part of the analysis. However, it is still valuable to be able to supplement the initial overall ANOVA F-testing, often with highest focus on the p-values with some good measures of overall effect size. In the widely used open source software PanelCheck (NOFIMA; ÅS, 2008) the inbuilt ANOVA results are visualized by multi-attribute bar plots of F-statistics combined with colour coding of the significance results. In this way the F-statistic is used as a kind of effect size measure. This can be a good approach, especially within PanelCheck, where the multi-attribute bar plot of the overall product differences are used only for single-factor product effects and with the same choice of F-test denominator across all the attributes of a plot.

However, the F-statistic itself is generally not the best measure of effect size as it depends on the number of observations for each product. And the various ANOVA mixed models, that we often use for such analysis also complicates the

relative effect size handling as generally in mixed models, different effects may have different noise structures, that is, different factors may be tested using different F-test denominators. Moreover, as was pointed out in Kuznetsova et al. (2015), it is important, specifically within the sensory and consumer field to be able to also handle more complicated settings than the most simple ones.

More recently, a number of new open source software tools with, among other things, focus on more extended type of mixed model ANOVA for sensory and consumer data have appeared. The ConsumerCheck (TOMIC et al., 2015), a tool developed in the same spirit as PanelCheck, offers quite general mixed model analysis of consumer data based on the newly developed more generic R-package lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a). In addition, in the still developing R-package SensMixed (KUZNETSOVA et al., 2015) and (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014b) one of the main purposes is to provide nice and visual multi-attribute interpretations of more complicated analysis. The resulting multi-attribute bar plots will then involve different factors with different number of levels and different number of observations within the levels. It may also involve different mixed model error terms for different factors. All of this calls for some careful thoughts on how to visualize the results of the (mixed) ANOVA results in the best possible way.

The purpose of the present study is to suggest better multi-attribute ANOVA plots for sensory and consumer data based on an effect size expressed in terms of relative pairwise comparisons. We will show how this has a close link to the Thurstonian d-prime, and as such is a generic measure that can be interpreted and compared across any attribute and situation. For balanced data settings, the measure is a simple transformation of an F-statistic making the approach easily applicable for anyone for these cases. For more challenging cases we offer a generic implementation of the method as part of the R-package SensMixed (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014b).

The paper is organized such that first, in Section 3.2, we introduce the basic notion of effect size (ES) in ANOVA framework and the concepts of d-prime. Then in Section 3.3, we define the effect size $\tilde{\delta}$. Next, in Section 3.4 it is shown how to estimate the $\tilde{\delta}$ ES measure for the standard mixed model with possible bias

correction. After this, in section 3.5 we illustrate the method on a multi-factorial sensory profile data set and compare the $\tilde{\delta}$ proposed here with the actual d-prime based on Thurstonian modelling. The paper ends with discussions in Section 3.6.

3.2 Cohen's d and d-prime - important effect size measures

Analysis of variance (ANOVA) is one of the most used and the most important methodologies when focus is on investigating product differences in sensory and consumer studies (NÆS; BROCKHOFF; TOMIC, 2010). ANOVA includes a particular form of null hypothesis statistical testing (NHST) used to identify and to quantify the factors that are responsible for the variability of the response. The null hypothesis for ANOVA is that the means of the factors are the same for all groups. The alternative hypothesis is that, at least, one mean is different from the others. An F-statistic is obtained in the ANOVA and the F distribution is used to calculate the p-value.

The NHST is a direct form and an easy way to conclude about the statistical significance of a factor, by considering a significance level and a p-value. However, it gets a lot of criticism from researchers of different fields. Yates (1951) observed that researchers paid undue attention to the results of the tests of significance and too little attention to the magnitudes of the effects, which they are estimating. NHST addresses whether observed effects stand out above sampling error by using a test statistic and its p-values, though it is not as useful for estimating the magnitude of these effects (CHOW, 1998).

As Sun, Pan and Wang (2010) observed, the fundamental problem with NHST is not that it is methodologically wrong; the misuse of NHST is the fault. Cohen (1994) pointed out that the NHST does not tell us what we want to know, and we so much want to know what we want to know, that, out of desperation, we nevertheless believe that it does!

The ongoing debate over statistical significance tests has resulted in alternative or supplemental methods for analysing and reporting data. One of the most frequent recommendations is to consider the effect size estimates to supplement p-values and to improve research interpretation (COE, 2002; COHEN, 1990, 1992,

1994; CUMMING; FINCH, 2005; DEVANEY, 2001; FAN, 2010; GRISSOM; KIM, 2012; KELLEY; PREACHER, 2012; STEIGER, 2004; SUN; PAN; WANG, 2010). Cohen (1990) affirms that the purpose should be to measure the magnitude of an effect rather than simply its statistical significance; thus, reporting and interpreting the effect size is crucial. Fan (2010) shows that p-value and effect size complement each other, but they do not substitute for each other. Therefore, researchers should consider both p-value and effect size.

Cohen (1992) established a relation between the effect size (ES) and NHST definitions: the ES corresponds to the degree in which the H_0 is false, i.e., it is a measure of the discrepancy between H_0 and H_1. Grissom and Kim (2012) states that whereas a test of statistical significance is traditionally used to provide evidence (attained $p-$value) that the null hypothesis is wrong; an ES measures the degree to which such a null hypothesis is wrong (if it is false).

In other words, an effect size is a name given to a family of indices that measure the magnitude of a treatment effect. It can be as simple as a mean, a percentage increase, a correlation; or it may be a standardized measure of a difference, a regression weight, or the percentage of variance accounted for. For a two-group setting, the ES quantifies the size of the difference between two groups, and may therefore be said to be a true measure of the significance of the difference (COE, 2002).

An important class of ES measures is defined by using the standardized effect size. In this class are included the Cohen's d, which is the difference measured in units of some relevant standard deviation (SD) (CUMMING; FINCH, 2005). Cohen's d is the ES index for the t test of the difference between independent means expressed in units of (i.e., divided by) the within-population standard deviation, which is given by:

$$d = \frac{\mu_a - \mu_b}{\sigma}$$

where μ_a and μ_b are independent means and σ is the within-population standard deviation.

There are several effect size measures to use in the context of an F-test for ANOVA. Cohen (1992) defined the effect size for one-way ANOVA as the standard deviation of the K population means divided by the common within-population

standard deviation:

$$f = \frac{\sigma_m}{\sigma} \tag{9}$$

where σ_m is the standard deviation of the K population means and σ is the within-population standard deviation.

A very similar measure of standard ES for ANOVA is the root-mean-square standardized effect (Ψ) presented by Steiger (2004). Considering the one-way, fixed-effects ANOVA, in which K means are compared for equality, and there are n observations per group the root-mean-square standardized effect is defined by

$$\Psi = \sqrt{\frac{\frac{1}{K-1}\sum_{i=1}^{K}(\mu_i - \mu)^2}{\sigma^2}} \tag{10}$$

where σ^2 is the mean square error. In fact, this could be just an interpretation of what the Cohen's f really is using $K - 1$ for expressing the standard deviation as opposed to using K as others might do. For the remainder of this paper we allow ourselves to consider the Ψ to be our version of the $Cohen's$ standardized ES measure for one-way ANOVA, such that for us "Cohen's $f = \Psi$".

The field of ES measures and estimation thereof is characterized by a certain level of confusion in the choice and use of the various ES measure names, where different names are used for almost the same measures. And, some names are used and defined for population versions of the measures whereas others for sample versions. In addition, the confusion is not diminished by the fact that many of these sample version measures will be biased estimates of the population versions, so often several alternative sample versions of the same population measure exist. It is not the aim of this paper to uncover and review this entire field. Rather we will be clear on exactly how we define the measures we use in both the population versions and the sample versions. Also, sometimes such ES measures are used for power and sample size computations in the planning phase, and at other times they are used for the actual data analysis. We will use it purely for data analysis interpretation.

Cohen's d for a two-sample setting has a close link with the Thurstonian d-prime, a signal-to-noise ratio from Signal Detection Theory (SDT), which is

widely used in sensory science. Mathematically speaking they are exactly the same: the difference between two means relative to a standard deviation. Only the contexts are usually different. The framework of SDT (GREEN; SWETS, 1966) and Thurstonian modelling (THURSTONE, 1927) make it possible to investigate the internal and external factors in sensory test and study how these factors influence subjects' test performance (HOUT, 2014).

The SDT approach was first introduced in sensory science with focus on exploring the factors influencing the perceptual process that integrates the information from the senses and the decision process (O'MAHONY, 1972, 1979), leading to more effective test designs (O'MAHONY, 1995). After this the focus shifted towards understanding and optimizing the decision processes in sensory tests leading to the development of more effective tests that are more predictive of consumer's reality (HAUTUS; O'MAHONY; LEE, 2008).

The Thurstonian approach is responsible for the biggest impact on the development of sensory difference discrimination test methods (e.g. the duo-trio, triangle, 2-AFC, 3-AFC). Since this kind of tests produce binary data, the statistical methods needed for analysing such data can be found among methods based on the binomial distribution and standard methods for analysing tables of counts. As pointed out by Ennis (1993) one of the weaknesses of working on the count scale is that it is test protocol dependent: the number of expected correct answers for the same products depend heavily on which test protocol that is carried out.

By transforming the number of correct answers into an estimate of the underlying (relative) sensory difference, the Thurstonian model gives the so-called d-prime (d'). The d-prime, which was defined to quantify the effect size, is the estimate of the size of a sensory difference from a particular test. This measure can be seen as generalized measure of sensory difference that expresses size of sensory differences. Since the d' is independent of the test method used, it 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 (HOUT, 2014). Although maybe the Thurstonian approach is most well-known for its use for sensory discrimination test protocols with binary or ordered categorical outcomes, it has also been suggested and used for ratings data, see e.g. Ennis

(1999) and Warnock, Shumaker and Delwiche (2006) and also in the context of multivariate analysis of ratings data as e.g. probabilistic multidimensional scaling, cf. (MACKAY; ZINNES, 1986). Brockhoff and Christensen (2010) and Christensen, Cleaver and Brockhoff (2011) showed how the Thurstonian approach in many cases could be viewed as and embedded into the so-called generalized linear model and/or ordinal regression theory and framework. One benefit of this is the ability to handle regression and ANOVA type analysis within the framework of a Thurstonian approach; where otherwise the most common Thurstonian approach would be to do repeated one- and two-sample computations on various subsets of the data.

3.3 Methods

We suggest using an ES measure that measures the average pairwise differences between the products or factor levels in question. More specifically, we define it as the root mean square of standardized pairwise differences, which in the balanced one-way ANOVA setting (I groups with n observations in each group), can be expressed as:

$$\tilde{\delta} = \sqrt{\frac{2}{I(I-1)} \sum_{i_1<i_2}^{I} \left(\frac{\mu_{i_1} - \mu_{i_2}}{\sigma} \right)^2} \tag{11}$$

where $\sum_{i_1<i_2}^{I}$ means the sum of all unique combinations of the two indices. The sum hence includes $I(I-1)/2$ terms, and it is clear that we have expressed the square root of the average of all standardized squared pairwise differences.

The first thing to notice is that the only difference to the Cohen's f or Ψ measures, defined above, is that products - usually in sensory and consumer applications the groups would represent different products - are compared pairwise rather than with the overall mean. This means that we have the following relation between $\tilde{\delta}$ and (our version of) Cohen's f in this balanced one-way ANOVA

setting:

$$\tilde{\delta} = \sqrt{2}\Psi = \sqrt{2}f$$

The formal (and short) proof of this is given in the Appendix A.

We need to use our $\tilde{\delta}$ ES measure also for multi-factorial settings as this will be an important part of the applications of this. Even though it may be more or less straightforward how this can be done, we believe that it is clarifying to at least express this formally in one of the simplest non-trivial extensions. For the replicated two-factor factorial design, the ANOVA model with main effects of A (α_i, $i = 1, \ldots, I$) and B (β_j, $j = 1, \ldots, J$) and interaction effects A×B (γ_{ij}), we define the $\tilde{\delta}$ ES measures as:

$$\tilde{\delta}_A = \sqrt{\frac{2}{I(I-1)} \sum_{i_1<i_2}^{I} \left(\frac{\alpha_{i_1} - \alpha_{i_2}}{\sigma} \right)^2} \tag{12}$$

$$\tilde{\delta}_B = \sqrt{\frac{2}{J(J-1)} \sum_{j_1<j_2}^{J} \left(\frac{\beta_{j_1} - \beta_{j_2}}{\sigma} \right)^2} \tag{13}$$

$$\tilde{\delta}_{A\times B} = \sqrt{\frac{2}{IJ(IJ-1)} \sum_{ij<i'j'}^{IJ} \left(\frac{\gamma_{ij} - \gamma_{i'j'}}{\sigma} \right)^2} \tag{14}$$

where the sum $\sum_{ij<i'j'}$ means all unique pairwise combinations of all IJ levels. Note, how this definition of the interaction ES measure is a “pure" interaction measure, where indeed all of the many combined levels are compared with each other, but only the real interaction effects are included, that is, the main effects have been removed from this measure. In this way, the size of the interaction ES measure is directly comparable with the size of the main ES measures.

Inspired by the 2-way interaction expression we can formulate a version of $\tilde{\delta}$ that would be applicable for any order of interaction effect $F = F_1 \times F_2 \times \cdots \times F_M$:

$$\tilde{\delta}_F = \sqrt{\frac{2}{K(K-1)} \sum_{k<k'} \left(\frac{\gamma_k - \gamma_{k'}}{\sigma} \right)^2} \tag{15}$$

where the sum $\sum_{k<k'}$ means the unique pairwise combination of all combinations of the levels of all the factors in $F = F_1 \times F_2 \times \cdots \times F_M$, and γ_k is the interaction effect for the k'th of all these combinations, where $k = 1, \ldots, K$ and K is the total number of combinations in the interaction effect. In addition, as above: the effects of all lower order effects have then been removed from the measure.

Finally, it is important to realize that these definitions also apply to situations where at the same time we are having yet other effects in the model including the possibility of these being regression (covariate) effects or any combination of such. With this is place we are now ready to begin the discussion of how to compute these measures in practice.

3.4 The sample estimation of the $\tilde{\delta}$ ES measures

3.4.1 The independent two- and multi-group one-way ANOVA case

In the two-independent-samples case with $n_1 = n_2 = n$, the absolute value of the pooled t-test statistic is:

$$|t| = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{2}s/\sqrt{n}} = \sqrt{\frac{n}{2}}\frac{|\bar{x}_1 - \bar{x}_2|}{s}$$

where $s = \sqrt{MSE}$ is the pooled standard deviation estimate. When there are only two means to compare, the t-test and the ANOVA F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$. So a simple rescaling of the root-F statistics will correspond to the "plug-in" sample version of $\tilde{\delta}$ in this case (as there is only one term in the sum that defines $\tilde{\delta}$):

$$\sqrt{\frac{2}{n}}\sqrt{F} = \frac{|\bar{x}_1 - \bar{x}_2|}{s}$$

Similarly for the balanced K-group one-way ANOVA setting, we can obtain a "plug-in"-sample estimate of $\tilde{\delta}$ by the same back transformation of the F-

statistic:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F}$$

This is almost directly clear from the definition of Ψ above.

3.4.2 Bias of sample estimates and possible bias corrections

The simple plug-in sample estimate that appeared in a natural way above is in fact not an unbiased estimate of the population $\tilde{\delta}$-value. And even though this may not necessarily prevent us from using such a plug-in approach for the visualization purposes, that are the main focus of the current paper, it is valuable to have some understanding of the bias mechanisms. Some way to possible bias corrections, at least in cases where this will be straightforward, would be valuable.

Assuming the standard normal based one-way ANOVA model, formally the F-statistic:

$$F = \frac{MS_{Product}}{MSE}$$

will have a non-central F-distribution as its sampling distribution. The mean of the F can then be found from basic probability and is well-known to be:

$$E(F) = \frac{nK}{nK-2}\left(\frac{n\sum_{i=1}^{K}(\mu_i-\mu)^2/(K-1)+\sigma^2}{\sigma^2}\right) = \frac{nK}{nK-2}\left(\frac{n}{2}\tilde{\delta}^2+1\right)$$

where we have re-expressed it in terms of our $\tilde{\delta}$ ES measure. We see that using the plugin sample estimate by the above given back transformation of the F-statistic will over-estimate the $\tilde{\delta}$ in two ways. Firstly, the fraction $\frac{nK}{nK-2}$ is always larger than 1. This bias mechanism comes from the fact that the mean of the fraction of two random variables is not the fraction of the means. Since in general for reasonably sized experiments, the number $\sqrt{\frac{nK}{nK-2}}$ will be rather small, this bias will most often not be important, and for most of what we do from here this will be ignored. But if wanted, a simple correction by the factor could be applied.

Secondly, and more importantly, we can see that the less biased back trans-

formation of the F-statistic would be to subtract 1 from it:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F-1}$$

The bias mechanism behind this effect comes from fact that when computing the variability between the sample means we also get some residual error as part of it, which is seen from the classical expected mean square expression for the expected value of the numerator of the F-statistic:

$$E(MS_{product}) = n\sum_{i=1}^{K}(\mu_i - \mu)^2/(K-1) + \sigma^2$$

As this bias can be non-trivial, and the smaller the F, the higher the relative bias, we recommend to correct for this whenever feasible.

3.4.3 Some standard mixed model sensory and consumer cases

For most sensory and consumer applications the proper model to use would be a mixed model of some kind, where at least effects related to assessors or consumers would be considered random, see e.g. Kuznetsova et al. (2015) and Næs, Brockhoff and Tomic (2010). Three such examples are the complete consumer preference study corresponding to a completely randomized block setting, the randomized replicated quantitative descriptive Sensory analysis (QDA) corresponding to a multi-attribute two-way (products-by-assessor) mixed ANOVA or the batched/sessioned replicated QDA corresponding to a three-way (products-by-assessor-by-batches) mixed ANOVA. These are the three cases that for single factor product study design and complete data can be handled by the PanelCheck tool, leading to either of the following three F-tests for product differences:

$$\begin{aligned}
F_{prod} &= \frac{MS_{prod}}{MSE} \\
F_{prod} &= \frac{MS_{prod}}{MS_{product\times assessor}} \\
F_{prod} &= \frac{MS_{prod}}{MS_{product\times assessor} + MS_{product\times session} - MSE}
\end{aligned}$$

Even though the significance statements in the latter two cases are based on the shown “mixed" error terms, the definition of $\tilde{\delta}$ can be interpreted as measuring the effects relative to the residual standard deviation. Clearly, there could be some potential additional interpretations of considering effect sizes in mixed models relative to other error components than the residual error, but we leave that for future research. In the current paper, we interpret effects relative to the best estimate of the average within individual and within-product variability, that is, the residual error estimate. This means that the back transformation of F-statistics only work for the first case. However there is an easy way to obtain bias corrected estimates of $\tilde{\delta}$ for the two other cases: simply run the fully fixed effect version of the models and then apply the back transformation on the product F-tests coming from these models:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{F_{FIXED} - 1} \tag{16}$$

Remember, that the fully fixed model should only be run to get the ES measure estimates - everything else, including the significance information, should be extracted from the proper mixed model.

3.4.4 Back transforming F-statistics more generally

The back transformation formula we have given only holds for balanced main effects. For balanced interaction effects, the proper more general bias-corrected back transformation formula becomes:

$$\hat{\tilde{\delta}} = \sqrt{\frac{2}{n}}\sqrt{\frac{DF}{K-1}}\sqrt{F - 1} \tag{17}$$

where n is the (same) number of observations for each level of the interaction, DF is the degrees of freedom for the interaction effect and K is the number of combined levels of the interaction factor. The proof is given in the Appendix A. These back transformations formulas can then be applied to any balanced situation for any main and interaction effect in cases where a fully fixed version of the relevant model is run.

It is possible to find back transformations formulas for other situations, e.g. covariate effects, non-balanced settings or even a general contrast effect. An alternative that can always be used is to simply extract the relevant effects from the model fit and then use the defining formula directly. In practice, this can be done e.g. by extracting so-called lsmeans and/or model parameter estimates from the model and use those in the defining formula. This is the approach used in the R-package SensMixed providing a method that works for any setting. The downside of this is the lack of bias correction in the estimates.

3.4.5 More general mixed models

For more general mixed models in sensory and consumer applications the product F-statistics can have a more complex form and the effects are estimated by complex weighted averages of the data making the approach of formulating a corresponding fully fixed model followed by a back transformation of a fixed F unfeasible as the effects could be differently estimated in the fixed model.

We suggest instead the general “plug-in" approach implemented in the Sensmixed package, Kuznetsova, Brockhoff and Christensen (2014b) for these situations - these work similarly for mixed models.

3.5 Examples

This section will contain an example to illustrate the method on a multi-factorial sensory profile data set. We also present a simple example to compare the $\tilde{\delta}$ with the actual d-prime calculations based on Thurstonian regression modelling. The analysis was performed using the SensMixed package. The R-code of the first example is given in the Appendix B. The TVbo data set are available in the SensMixed package.

3.5.1 Example 1: Multi-way product structures in sensory profile data

The TVbo data set comes from the high-end HIFI company Bang & Olufsen A/S, Struer, Denmark. The main purpose was to test products, specified by two features: Picture (factor with four levels) and TVset (factor with three levels). The 12 combinations of TVset and Picture were assessed by a sensory panel composed by eight trained panelists for a list of 15 different response variables (characteristics of the product) in two replications. The data is available in the SensMixed package named TVbo.

To specify the mixed model, the main effect Assessor plus interactions between Assessor and product effects (TVset and Picture and the interaction TVset:Picture) are considered random effects. The fixed part contains a multi-way product structure: two main effect TVset and Picture and an interaction between them. The 15 attributes (Color Saturation, Colour Balance, Noise, Depth, Sharpness, Light Level, Contrast, Sharpness Movement, Flickering Stationary, Flickering Movement, Distortion, Div Glass Effect, Cutting, Flossy Edges, Elastic Effect) can be analysed all together using the SensMixed package.

In Figure 1 a multi-attribute bar plot based on the F-values from the mixed model is presented combined with colour coding of the significance results. Since the mixed model specified here has three fixed effects (TVset, Picture and interaction), the F-tests have different mixed model error term for each effect. In this way the F-statistic is not comparable because the F-test denominators are different across the attributes.

Looking into the multi product structure given by Figure 1 we can see that the main effect TVset is significant for 13 of 15 attributes; the main effect Picture and the interaction are significant for 11 of 15 attributes. For the attributes 2, 4 and 13 for instance, the main effect TVset is significant and Picture is not significant. It means that, for these attributes the products differ mostly due to the effect of TVset. In that way, for the attribute 8 the products differ mostly due to Picture. For the attribute 10, all fixed effects are not significant, that means the assessors were not able to discriminate the products for this attribute. For the remainder attributes, 1, 3, 5, 6, 7, 11, 12, 14 and 15, both main effects are significant and also

the interaction, except for the attribute 12. Since the number of levels of the two main effects are different, the F test are not comparable.

In Figure 2 is presented the alternative bar plot to visualize the (mixed) ANOVA results based on the $\hat{\tilde{\delta}}$, the effect size measure obtained from the back transformation on the product F-tests coming from the fixed model for TVbo data. Comparing the bars of the delta-tilde plot (Figure 2), it can be seen that the effect of TVset is stronger than the effect of Picture for the attributes 1, 2, 4, 5, 6, 7 and 13. The effect of TVset for attribute 6, for instance, is much stronger than all the effects for the other attributes. The effect of Picture is stronger for the attributes 3, 8, 9, 11, 12, 14 and 15 than for the other attributes. The effect of interaction is stronger for the attribute 11 than for the other attributes. It is important to note that the Figure 2 gives us relevant information regarding the size of each effect. Furthermore the plot presented in Figure 2, based on the delta-tilde estimates, makes the bar heights better comparable, especially when the effects have different number of levels.

Even when the levels of the effects are the same, the delta-tilde plot can be a better visual tool, especially when there are a F-statistics much larger than the others, e.g. the F-statistics for the TVset effect for the attributes 6 and 15 given in Figure 1. It makes the small values difficult to visualize. With the back transformation, the ES estimates presented in the Figure 2 has a much smaller range which makes the bar heights better comparable.

For extensions of such multi-attribute bar plots into more complex models, a similar transformation is suggested and becomes more important as the transformation depends on the number of observations within factor levels, and hence makes bar heights better comparable for factors with differences in number of levels.

Now let us look more closely into an attribute to see how the $\hat{\tilde{\delta}}$ was calculated for each effect. Considering the Attribute 7 as an example. To obtain the bias corrected estimates of $\tilde{\delta}$ we first run the fully fixed effect version of the model and then we apply the back transformation on the product F-test from this model, cf. Table 1. It is important to keep in mind that the fixed effect model is used only to get the product F-tests to apply the back transformation to obtain the ES measure

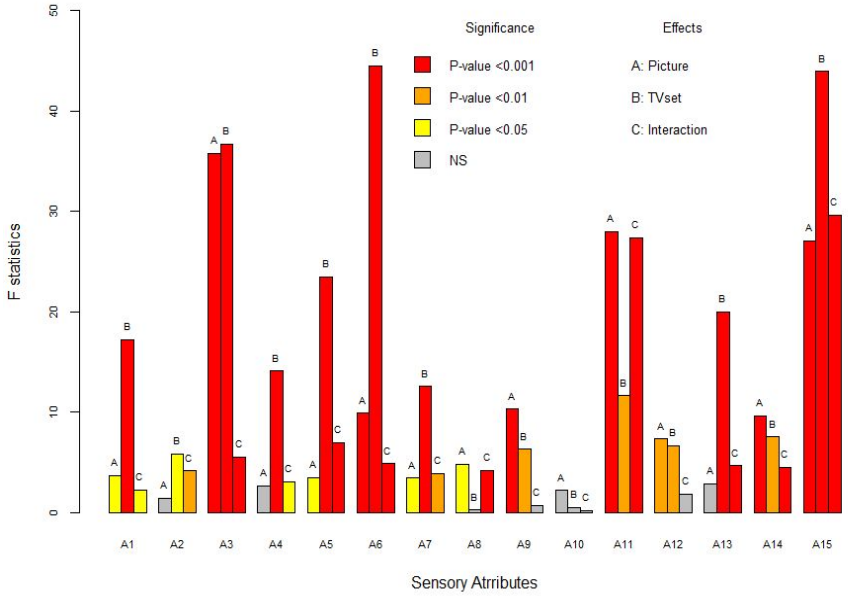

Figure 1 Bar plot for F values for fixed effects of TVbo data.

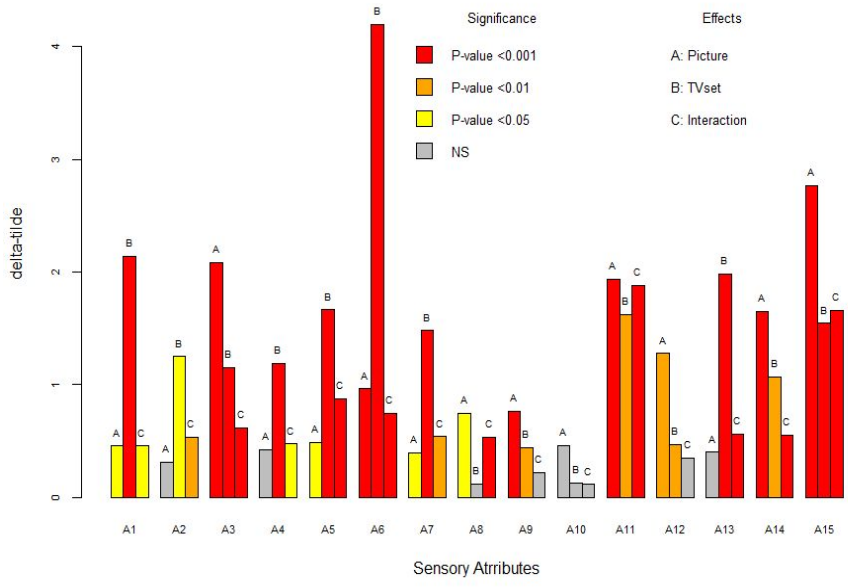

Figure 2 Bar plot based on delta-tilde for fixed effects of TVbo data.

estimates. The significance information should be extracted from the proper mixed model.

Table 1 ANOVA table for the fixed effect model for attribute 7 of TVbo data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TVset	2	247.41	123.70	70.01	0.0000
Picture	3	19.84	6.61	3.74	0.0136
TVset:Picture	6	44.98	7.50	4.24	0.0007
Assessor	7	130.87	18.70	10.58	0.0000
TVset:Assessor	14	137.93	9.85	5.58	0.0000
Picture:Assessor	21	22.51	1.07	0.60	0.9047
TVset:Picture:Assessor	42	99.83	2.38	1.35	0.1183
Residuals	96	169.64	1.77		

The back transformation of the F-statistics for the main effect is calculated according to the formula (16):

$$\hat{\tilde{\delta}}_{TV} = \sqrt{\frac{2}{64}}\sqrt{70.0045 - 1} = 1.47$$

$$\hat{\tilde{\delta}}_{Picture} = \sqrt{\frac{2}{48}}\sqrt{3.7421 - 1} = 0.34$$

For the interaction effect, we get the back transformation of the F-statistics from the more general bias-corrected transformation given by the formula (17):

$$\hat{\tilde{\delta}}_{TV*Picture} = \sqrt{\frac{2}{16}}\sqrt{\frac{6}{11}}\sqrt{4.2422 - 1} = 0.47$$

The delta-tilde estimates for the product effects (TVset, Picture and the interaction TVset:Picture) for the attribute 7 is presented in Figure 3. Since the delta-tilde estimates represents the effect size, the heights of the bars can be comparable between each other. From the Figure 3 we can see that the delta-tilde estimate for TVset is much larger than the others, which means that the effect of TVset is stronger than the effect of Picture for the attribute 7. So the impact of TVset effect on the ability to discriminate between products is higher than the im-

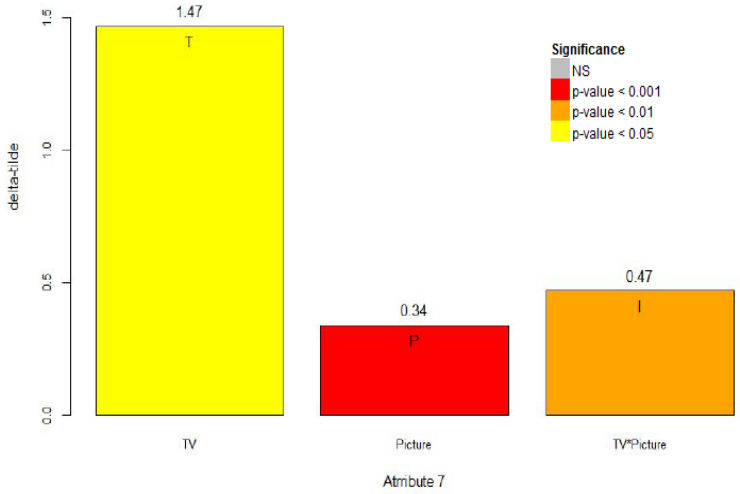

Figure 3 Bar plot for delta-tilde based on F-statistics from fixed effects model for attribute 7.

pact of Picture. When interpreting the $\tilde{\delta}$-values we must remember that these are expressing average pairwise differences. This means that if there is only a single product that differs from the rest, say, and the remaining ones are really the same, it will tend to appear as a small average effect in the plot - but potentially still statistically significant. These plots cannot substitute a good *post hoc* analysis of product differences.

3.5.2 Example 2: Comparison with d-prime from Thurstonian model - simple example

To compare the $\tilde{\delta}$ with the d-prime from Thurstonian model we will use the simplest example considering a subset of the TVbo data. Taking the average of TVset1 and TVset2 by Picture for the 8 Assessors we get the subset described in the table 2. Table 3 gives the ANOVA table for the subset of TVbo data.

The ES measure estimates for this situation is given by the difference be-

Table 2 Subset of TVbo data

Assessor	1	2	3	4	5	6	7	8	Mean
TVset1	3.7	5.5	7.1	2.6	11.3	9.1	8.8	8.3	7.0500
TVset2	1.4	4.0	4.8	2.1	7.4	5.2	2.1	4.1	3.8875

Table 3 ANOVA table for subset of TVbo data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tvset	1	40.01	40.01	6.38	0.0243
Residuals	14	87.85	6.27		

tween the independent means divided by residual error estimate.

$$\hat{\tilde{\delta}} = \frac{7.05 - 3.8875}{\sqrt{6.27}} = 1.26$$

To calculate the "real" d-prime from Thurstonian model we use the ordinal package (CHRISTENSEN, 2014). First the subset presented in the table 2 are categorized from 1 to 10, since the response in the cumulative link model (CLM) is usually interpreted as an ordinal response with levels ordered. The categorized data is presented in table 4. Then we obtain the d-prime from the cumulative link model function (see Appendix B) which is equal to 1.26.

Table 4 Categorized data for subset of TVbo data

Assessor	1	2	3	4	5	6	7	8
TVset1	3	5	6	2	10	8	8	8
TVset2	1	3	4	2	7	5	2	4

We can see that the close link between delta-tilde, the effect size in an ANOVA framework, and the Thurstonian d-prime, discussed in the section 3.2, can be confirmed by a comparison between the real d-prime calculation and the delta-tilde estimate.

3.6 Discussion

In this paper we have suggested the use of ES measures as a visual tool to improve the interpretation of the ANOVA table in Analysis of Variance. In spite of having been discussed in literature for decades, ES measures have not been used extensively for this purpose. Instead, more focus has been on the *post hoc* part of the ANOVA data analysis. We believe that even though the ES plots suggested here cannot substitute a good *post hoc* analysis, they are valuable additional tools for a good and relevant interpretation of the ANOVA table, and can help to move the focus a bit away from purely looking at p-values but rather focusing on the size of the effects (but still using the p-value information). And this becomes particularly useful in situations with more than a single factor and with several attributes.

It could be a relevant next step to work in the development of significance statements and effect confidence intervals. For now, we suggest a simple transformation on the F-statistics from ANOVA to obtain the ES measures. We also mentioned that one could pursue various explicit extended versions of how to back transform F-statistics to give the $\tilde{\delta}$-measure, but here we just use the simple transformation with the purpose of improving the visual interpretation. The approach transforms the bar plots of F-statistics by re-scaling the bar heights and gives the average pairwise d-tilde between products. This has the same interpretation as the real d-prime calculation from Thurstonian approach.

Acknowledgements

This work is a part of the Senswell project funded by Innovation Fund Denmark (grant no. 0603-00418B) and the ConsumerCheck project funded by The Danish AgriFish Agency under the Innovation Law (grant no. 3414-08-02347) and the Science without Borders program funded by the Brazilian government agencies CAPES and CNPq.

4 TOOLS FOR MIXED MODELLING OF SENSORY DATA

4.1 Introduction

In the half of last century many scientists had dreamed of the day when there would be possible to improve their statistical analysis by using computers. We would be glad to realize that we are living in these days. Actually, the analysis of data using softwares have brought enormous contribution for the science. Among the commercial and open source softwares for statistical analysis, the free available software R plays an important role. With numerous packages developed by the advanced users, the software R has become the most powerful and most widely used statistical software nowadays. Due to its elegance and power in academy, the software R has exploded in popularity and functionality, emerging as the data scientist's tool of choice. The sensometric scientists have also embraced R to solve their most challenging problems in fields ranging from mixed modelling analysis of variance to more complex as the Bradley-Terry and Thurstonian models, as well as multivariate methods, as Correspondence Analysis (CA), Multiple Correspondence Analysis (MCA), Principal Component Analysis (PCA) and Multiple Factor Analysis (MFA). The result has been many R packages developed specially for sensory science data such as sensR (CHRISTENSEN; BROCKHOFF, 2015), SensoMineR (HUSSON; LE; CADORET, 2014), lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a) and SensMixed (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014b).

In this Chapter, we present a review about the R packages lmerTest and SensMixed, both developed for mixed modelling of sensory data. We will focus more on the second one, which has an implementation of the delta-tilde method described in Chapter 3.

Another open source tool, which deserves attention, is the user-friendly software PanelCheck (NOFIMA; ÅS, 2008), widely used for high throughput analysis of sensory quantitative descriptive analysis (QDA) data (AMORIM et al., 2014). This software includes visual tools for simple linear mixed model (DAHL; TOMIC; NÆS, 2008) and due to its simplicity in use, it became very popular among the

sensory practitioners. An example of a mixed model for sensory study using the tools will be presented in this Chapter.

4.2 Overview of the recently developed tools for fitting mixed models to sensory data

Mixed effects models are used as an appropriate choice for analysing sensory and consumer data. In order to extract important attribute-wise product difference information, the analysis of variance (ANOVA) methods are generally applied (LAWLESS; HEYMANN, 2010; NÆS; BROCKHOFF; TOMIC, 2010). There are several commercial and open source softwares for fitting mixed models to sensory and consumer data. Still applying such models may be challenging for a sensory practitioner. The challenges arise as to which model to consider, which effects should be chosen as random and what are the interpretations of the results (KUZNETSOVA et al., 2015). In this Chapter we focus on the most recently tools developed for helping sensory practitioners to apply mixed models for sensory and consumer data. This include the PanelCheck software (NOFIMA; ÅS, 2008) and the two R packages lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a) and SensMixed (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014b).

4.2.1 PanelCheck software

PanelCheck, an open source software, was developed in collaboration between Nofima (Norwegian Institute of Food, Fisheries and Aquaculture Research) and the Technical University of Denmark (DTU) for analysis of sensory data (NOFIMA; ÅS, 2008). It includes univariate and multivariate methods, which provide plots for both panel performance monitoring and analysis of product difference. By visualizing different type of information in a set of various plots, the panel leader can investigate the performance of individual assessors and can detect individual differences among assessors. PanelCheck also includes visual tools for simple mixed models for multi-attribute data. By using graphical methods, PanelCheck provides an easy approach for the interpretation of results of sensory

data.

PanelCheck has a graphical user interface (GUI) and due to the simplicity in use, it is now used extensively. However, as pointed out by Amorim et al. (2014) the scope of the mixed modelling in PanelCheck is limited in several ways. For instance, it cannot handle with 2-(or higher) way product structure. Furthermore, PanelCheck works just for situations where the data are balanced. In sensory and consumer data is common situations where we deal with missing values due to one or more assessors who did not complete all replication. The software PanelCheck can still be a valuable tool to provide relevant ANOVA information if we consider products as single factor and use missing values imputation or exclude the assessors with missing values. Kuznetsova et al. (2015) pointed out many situations that really call for a more complex analysis, for instance:

- Unbalanced sensory profile data (due, for example, to missing observations).
- Incomplete consumer preference data.
- 2-(or higher) way product structure in sensory profile data.
- 2-(or higher) way product structure in consumer preference data (Conjoint analysis).
- Extending Conjoint analysis to include consumer background/design variable or factors/covariates.
- Complex blocking, product, replication, product batch structures in as well sensory as consumer preference data.
- Extending external preference mapping to include product and consumer background/design variables factors/covariates.

The R package lmerTest was developed by Kuznetsova, Brockhoff and Christensen (2014) in order to help to answer the questions above.

4.2.2 lmerTest package

In order to gain in exibility, the software R can be used. The main R package, which proposed functions that allows integrating mixed models is the lme4 package (BATES et al., 2014). Even with all advantages from lme4 package, which make it be the most used R package for fitting mixed effect models, it comes with one drawback: it does not provide every useful results (KUZNETSOVA et al., 2015). For instance, it does not provide p-values associated with parameters estimates and models terms based on F-statistics in ANOVA table.

The lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a) package builds on top of the lme4 and extends it performing different kinds of tests on lmer objects. An ANOVA table with corrected F-tests of type III hypotheses for fixed effect terms and parameters using Satterthwaite or Kenward-Roger approximations to the denominator degrees of freedom is provide by the lmerTest package (KUZNETSOVA et al., 2015). The package also provides the log-likelihood ratio tests for the random part of the model with one degree of freedom, which means, testing one effect in a time. As *post hoc* analysis, the least square means (population means) and differences of least squares means for the factors of the fixed part are presented with corresponding plots (KUZNETSOVA et al., 2015).

To use the lmerTest package, it is necessary to install and to load the package by typing in the R console the following lines:

```
install.packages("lmerTest")
library(lmerTest)
```

One of the purpose of lmerTest package is to investigate the mixed model, in an automated way, and incorporate the necessary random-effects by sequentially removing non-significant random terms in the model, and similarly test and remove fixed effects (KUZNETSOVA et al., 2015). The step function can be used to perform this automated complex mixed model selection. The step function investigates the random and fixed terms in the mixed model performing a backward elimination of non-significant effects, starting from the random effects, and then the fixed ones (KUZNETSOVA et al., 2015). As a result, the step function gives a list of effects that should be kept in the model in terms of significance and gives the

best, by principle of parsimony, model together with *post hoc* analysis presented in tables and plots. According to Kuznetsova et al. (2015) the automated model selection involves three steps:

1. Specification of the model (by the user).
2. Simplification of the random effects structure.
3. Simplification of the fixed effects structure.

Step 1: Specification of the model

First, the user needs to construct the initial mixed effects model. It is necessary to specify the initial (maximal) model where the fixed and random parts contain all explanatory variables and as many interactions as possible. To specify the mixed effect model the lmer function from lme4 package is used. The arguments to a lmer call are as follow:

```
lmer(formula, data=NULL, REML=TRUE,
    control=lmerControl(), start=NULL, verbose=0L,
    subset, weights, na.action, offset, contrasts=NULL,
    devFunOnly=FALSE, ...)
```

The user can specify the formula of the model using the lmer syntaxes and call the data set. The mixed model formula can be specified to lmer function as, e.g.:

```
response variable ~ product + (1|assessor)
                    + (1|product:assessor)
```

where `product` specify a fixed effect, `(1|assessor)` is a random effect and `(1|product:assessor)` is the random interaction between the fixed and the random effects. To specify the interactions we have two basic variants:

- `a:b` for an interaction between `a` and `b` effects
- `a*b` which expands to `a + b + a:b`

For random factors, you have three basic variants:

- Intercepts only by random factor:
  `(1|random.factor)`
- Slopes only by random factor:
  `(0 + fixed.factor|random.factor)`
- Intercepts and slopes by random factor:
  `(1 + fixed.factor|random.factor)`

Note that variant 3 has the slope and the intercept calculated in the same grouping, i.e. at the same time. If we want the slope and the intercept calculated independently, i.e. without any assumed correlation between the two, we need a fourth variant:

- Intercept and slope, separately, by random factor:
  `(1|random.factor) + (0 + fixed.factor|random.factor).`

Step 2: Analysis of the random effects

The `step` function will do the automated simplification of the random part and will present a table with the order in which one effect has been eliminated from the initial model. Therefore, let `M` be the mixed model from the first step, for example. To do elimination process the following R code can be used:

```
M <- lmer(attribute ~ Product +
                      (1|Assessor) +
                      (1|Assessor:Product) +
                      (1|Repeat) +
                      (1|Repeat:Product),
                      data = dataset)
s <- step(M)
```

For the simplification of the random effects done by the step function, the p-values are based on likelihood ratio tests. The p-values for each random effect are calculated and the one with the highest p-value that is less than the significance level α is eliminated and a new model is constructed without this effect.

The loop stops when there are no more non-significant effects or when there are no more random effects to be tested (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a). The table with the simplification of the random part of the model can be extracted with the command:

```
s$rand.table
```

If the reduction of the random part is not required, just specify it including the argument:

```
reduce.random = FALSE
```

Step 3: Analysis of the fixed effects

The `step` function will also do the automated analysis for fixed effects. In the fixed effect elimination process, the p-values are calculated from F test based on Satterthwaite's (default) or Kenward-Roger approximation. The p-values for each fixed effect are calculated and the one with the highest p-value that is less than the significance level α is eliminated and a new model is constructed without this effect. The loop stops when there are no more non-significant effects or when there are no more effects to be fixed tested. The table with the simplification of the fixed part of the model can be extracted with the command:

```
s$anova.table
```

The step function performs the backward elimination of the random part following by the backward elimination of the fixed part. After that the LSMEANS and differences of LSMEANS for the fixed part of the model are calculated. The commands bellow extracts the LSMEANS with p-values and confident intervals and the differences of LSMEANS with p-values and confident intervals respectively.

```
s$lsmeans.table
s$diffs.lsmeans.table
```

The final model by the principle of the parsimony (KUZNETSOVA et al., 2015) can be extracted using the command:

```
s$model
```

The plots and *post hoc* analysis can be obtained with the command:

```
plot(s)
```

4.2.3 SensMixed package

Kuznetsova et al. (2015) gave a one-step forward in facilitating analysis of sensory data in complex situations with the lmerTest and extended it developing the SensMixed package (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014b). This new package contains tools for advanced statistical methods within a mixed effects framework using the same technique of the automated analysis as in Kuznetsova, Brockhoff and Christensen (2014a) but applied simultaneously to all attributes and presenting the results in a compact and efficient way (KUZNETSOVA et al., 2015). The SensMixed package provides results of the analysis of random and fixed effects presented in tables and plots, including the new delta-tilde plot, present in Chapter 3. Beyond that, the SensMixed provides a tool to analyse the extended versions of the Mixed Assessor Model (MAM), a model that corrects for a possible scaling effect (BROCKHOFF; SCHLICH; SKOVGAARD, 2015). In this extended version of the MAM, a possible multi-way product structure can be accounted for together with the 3-way error structure, where a replicate effect is also accounted for (KUZNETSOVA et al., 2015).

The SensMixed package provides an intuitive and easy-to-use graphical interface implemented via the R package named shiny (CHANG et al., 2015). Apart from providing this easy-to-use interface for advanced statistical methods within a mixed effects framework, the application includes such crucial functionalities as importing the data in different formats, presenting results in tables and plots as well as saving them. This allows for efficient analysis of sensory data and enables the sensory practitioner and non-statistician to focus on results of the statistical analysis rather than spending time on trying to apply algorithms on the data by themselves (KUZNETSOVA et al., 2015). Together with its application, all that makes the SensMixed package a very valuable tool for sensory practitioners as it requires no skills in R-programming and provides advanced statistical methods

for analysing sensory and consumer data.

In order to run the application, one needs to install the SensMixed package and call the `SensMixedUI` function by typing in the R console the following lines:

```
install.packages("SensMixed")
library(SensMixed)
SensMixedUI()
```

A number of modelling options that allow to easily constructing and analysing in a proper manner a broad range of complex mixed effects models are provided. These options make the model building more flexible and advanced (KUZNETSOVA et al., 2015). The results of the analysis are visualized in various plots and tables, helping sensory practitioners to visually detect performance issues without having to know all details on the statistical methods. The main modelling controls of SensMixed are described by Kuznetsova et al. (2015) as the following ones:

- **error structure**

  No Rep: assessor effect and all possible interactions between assessors and product effects

  2-WAY: **No Rep**, replicate effect and interaction between assessor and replicate effects

  3-WAY: assessor and replicate effect and interaction between them and interaction between them and Product effects

- **product structure**

  1 main product effects

  2 main product effects and 2-way interactions between them

  3 main product effects and all possible interactions between them

- **scaling correction**

  Yes

No

These controls are responsible for the specification of the mixed effects model. **error structure** stands for the specification of the random part of a mixed effects model. **error structure = 3-WAY** produces the maximal possible random structure. This option is advised in Kuznetsova et al. (2015). However if, for example, from the studies it is known that there is no replication effect, then the **No-Rep** option can be considered. If it is known that there is no interaction between replication and product effects, then the **2-WAY** option may be chosen, which also conducts the analysis in a faster way (KUZNETSOVA et al., 2015).

The **product structure** is responsible for specification of the fixed part of the mixed effects model. If there is no multi-way product structure in the data, then all options produce the same fixed part. Otherwise, the option **3** produces the maximal possible fixed structure (KUZNETSOVA et al., 2015).

If one chooses to correct for scaling, then the Mixed Assessor Model is constructed. According to Brockhoff, Schlich and Skovgaard (2015) whenever the scaling is significant it is advisable to correct for it, since the tests for the product effects become more powerful.

According to the specified modelling controls the mixed effects models are constructed for all attributes using the lme4 package (BATES et al., 2014) and then the `step` method of the lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a) is applied to each model. In all cases, the fixed part is not simplified. By default the non-significant random effects are eliminated from the model according to the specified by a user Type 1 error (0.1 the default one). However, one may require not to eliminating the random effects, or specifying which effects should be kept in the model even if not being significant (KUZNETSOVA et al., 2015).

4.3 Example: mixed model analysis of sensory study

In this section, we present a mixed model analysis of a sensory study using the software PanelCheck and the R packages lmerTest and SensMixed. The aim is to show the functionality of these tools and help the sensory practitioners to apply advanced statistical techniques to get important information about their

studies.

4.3.1 Sensory study of car audio system

The data presented here comes from the company Bang and Olufsen A/S, Struer, Denmark. The purpose of this study was to rate products, specified by three features: Car (sound system), SPL (reproduction of sound pressure level) and Track (music program). The trained audio panel was composed by 10 assessors (Participant) who evaluate 90 products (CLIP) for 8 different response variables (Attributes) in 2 replications. Only 8 assessors completed both replications. Here is a brief description of the car audio system data:
CLIP: factor with 90 levels specified by:

- Car: factor with 6 levels
- SPL: factor with 3 levels
- Track: factor with 5 levels

Participant: factor with 10 levels
Replicate: factor with 2 levels
Attributes: response variables (8)

The names of the 8 attributes were translated from Danish [4] as: continuous noise, accuracy in the lower frequency range, accuracy in the upper frequency range, reverberation, stereo effect, strength of the bass range, strength of the treble range and strength of the mid-range. For simplicity we will call them att1, att2, ..., att8 and the dataset will be called SoundBO data.

4.3.2 One-way product analysis using PanelCheck

The modelling in PanelCheck consider product as fixed effect and both assessor and assessor-by-product interaction as random effects, getting simple linear mixed model. Due to its limitation, PanelCheck cannot handle with 2-(or higher)

[4] Kontinuerligstøj, Præcisioninedreområde, Præcisioniøvreområde, Rumklang, Stereovirkning, Styrkenafbas, Styrkenafdiskant og Styrkenafmellemtone

way product structure such as SoundBO data. Beyond that, the data must be complete and balanced. So an approach to analyse such data could be considering one product factor with 90 levels for the 2-way ANOVA mixed model with the assessor-by-product interaction as the error structure (LAWLESS; HEYMANN, 2010) and removing the assessors who did not complete both replications.

To analyse SoundBO data using PanelCheck, the assessors 1 and 6 were removed from the data set to consider the data 100% complete and balanced (no missing value). We also consider the 90 products (CLIP) formed by 6-by-3-by-5 combinations as a single factor since in PanelCheck, the model cannot take into account this 3-way product structure. The simple one-way product mixed model for one attribute y_{ijk} can be specify as:

$$y_{ijk} = \mu + a_i + \rho_j + r_k + b_{ij} + c_{ik} + d_{jk} + \epsilon_{ijkl} \tag{18}$$

$$\begin{aligned}
a_i &\sim N(0, \sigma^2_{assessor}),\\
r_k &\sim N(0, \sigma^2_{replicate}),\\
b_{ij} &\sim N(0, \sigma^2_{assessor \times CLIP}),\\
c_{ik} &\sim N(0, \sigma^2_{assessor \times replicate}),\\
d_{jk} &\sim N(0, \sigma^2_{replicate \times CLIP}),\\
\epsilon_{ijkl} &\sim N(0, \sigma^2_{error}).
\end{aligned}$$

where y_{ijk} corresponds to the attribute in study. The Greek letters represent fixed effects and the Latin letters represent the random effects. The main fixed effect of product (CLIP) is represented by ρ_j, $j = 1, 2, ..., J$. The random part of the model is compounded of the main effect Assessor, represented by a_i, $i = 1, 2, ..., I$; the main effect replicate, represented by r_k, $k = 1, 2, ..., K$ plus the interaction between Assessor and CLIP, represented by b_{ij}; the interaction between Assessor and replicate represented by c_{ik} and the interaction between replicate and CLIP represented by d_{jk}. The hypothesis of interest in model (18) is that there is no average product difference.

The results of the one-way product analysis using PanelCheck present a

barplot for each effect across all attributes (Figure 4). The colour of the bars (yellow, orange, red) indicate that there is a significant effect and a gray bar means that there is not. The bar size is equal to the F-statistic for each effect and the colour of the bar indicates the corresponding p-value: Yellow: $0.01 < p$-value < 0.05, Orange: $0.001 < p$-value < 0.01 and Red: p-value < 0.001.

From Figure 4 it can be seen that replicate effect is non-significant for all attributes; the interaction between replicate and CLIP is significant only for the attribute 1 ($0.01 < p$-value < 0.05); however there is a significant interaction between assessors and replicate for all attributes, except for the attribute 3. The product effect (CLIP) was highly significant ($p < 0.001$) for all attributes. That means the assessors has been able to discriminate between the products; therefore, we do not exclude any attributes from further analysis.

The product pairwise comparisons may be extracted, although, since there are 90 levels in the CLIP factor, there are $C_{90}^2 = 4005$ pairwise comparisons, which are indeed hard to interpret. Hence, considering a multi-way product structure might simplify the analysis of product differences by comparing product features plus some additional insight into the data might be gained. In such cases, lmerTest and SensMixed packages can be used to analyse the data considering the three effects Car, Track and SPL separately.

4.3.3 3-way product analysis using lmerTest

In the model (18), the combinations of Car, SPL and Track features form the 90 products. In order to consider a multi-way product structure the three main effects and all possible interactions between them should be consider instead of one CLIP effect. According with Kuznetsova et al. (2015) it gives more insight into the data. By using the lmerTest package, it is possible to fit a model that account for that multi-way product structure. Furthermore, the ten assessors can be considered in the analysis, since the lmerTest can deal with missing values.

To analyse data using lmerTest, the first step is to construct the maximal possible model according to (KUZNETSOVA et al., 2015). From the one-way product analysis in the previous section, we have deduced that Replicate effect and the

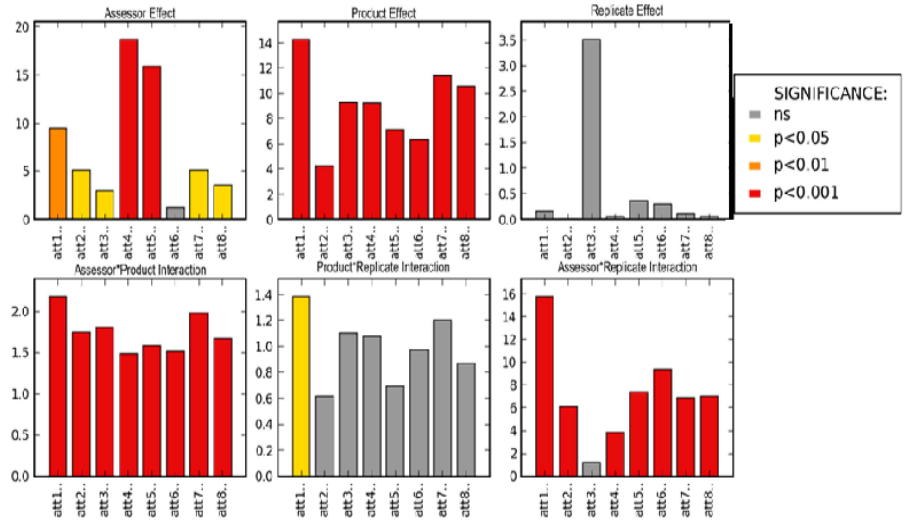

Figure 4 PanelCheck plot: F statistics from 2-way ANOVA for Sound data.

interaction between Replicate are non-significant for all attributes and CLIP are non-significant for all attributes, except for the attribute 1. However, there is a significant interaction between assessors and replicate for all attributes, except for the attribute 3. The non-significant effects identified in the PanelCheck analysis does not need to enter in initial the model.

To construct the initial model for SoundBO data set `assessor` and `replicate` are considered as random effects. The three main effects that constitute the product structure (Car, SPL and Track) are considered as fixed effects. The initial linear mixed model for one attribute y_{ijklm}, would be given by:

$$\begin{aligned} y_{ijklm} =& \mu + \delta_j + \lambda_k + \theta_l + \tau_{jk} + \nu_{jl} + \rho_{kl} + \gamma_{jkl} + \\ & a_i + r_m + b_{ij} + c_{ik} + d_{il} + p_{im} + \\ & e_{ijk} + f_{ijl} + g_{ikl} + h_{ijkl} + \epsilon_{ijklm} \end{aligned} \tag{19}$$

$$
\begin{aligned}
a_i &\sim N(0, \sigma^2_{assessor}),\\
r_m &\sim N(0, \sigma^2_{replication}),\\
b_{ij} &\sim N(0, \sigma^2_{assessor\times Car}),\\
c_{ik} &\sim N(0, \sigma^2_{assessor\times Track}),\\
d_{il} &\sim N(0, \sigma^2_{assessor\times SPL}),\\
p_{im} &\sim N(0, \sigma^2_{assessor\times replication}),\\
e_{ijk} &\sim N(0, \sigma^2_{assessor\times Car\times Track}),\\
f_{ijl} &\sim N(0, \sigma^2_{assessor\times Car\times SPL}),\\
g_{ikl} &\sim N(0, \sigma^2_{assessor\times Track\times SPL}),\\
h_{ijkl} &\sim N(0, \sigma^2_{assessor\times Track\times SPL\times Track}),\\
\epsilon_{ijkl} &\sim N(0, \sigma^2_{error}).
\end{aligned}
$$

where y_{ijklm} corresponds to the attribute in study and the Greek letters represent fixed effects and Latin letters represent random effects. The main fixed effects Car, Track and SPL are represented by δ_j, λ_k, θ_l accordingly; the two-way interactions between the fixed effect are represented by τ_{jk}, ν_{jl} and ρ_{kl}; and the three-way interaction is represented by γ_{jkl}. The random part of the model is compounded of the main effect Assessor represented by a_i plus the interactions between Assessor and the fixed part of the model; and the main effect replicate represented by r_m and the interactions between replicate and Assessor.

The model (19) has a quite complex error structure. We observe that 10 random effects form the random part of the model. It might be that not all of these effects contribute to the systematic variation in the data and therefore could be excluded from the model (KUZNETSOVA et al., 2015). The `step` method from the lmerTest package is used, that finds a parsimonious random structure by sequentially removing non-significant random effects (KUZNETSOVA et al., 2015).

To specify the fixed part of the model (19) in lmerTest, the lmer syntaxes can be used. `Car*SPL*Track` represents the three main fixed effect (Car, SPL and Track) and all possible interactions for them. The lmer syntaxes for the random effects will be `(1|Part)` and `(1|Rep)`. Finally we have to specify the random interactions. In lmerTest as in lme4, the users must to specify one model

for each attribute. The command line in lmer to specify the mixed model for the attribute 4 that corresponds to the mixed model (19) is given as an illustration:

```
M1 <- lmer(Att4 ~ Track*Car*SPL+
           (1|Part) + (1|Rep) + (1|Rep:Part) +
           (1|Part:Track)+(1|Part:Car)+(1|Part:SPL) +
           (1|Part:Track:Car)+(1|Part:Track:SPL) +
(1|Part:Car:SPL)+(1|Part:Car:SPL:Track),
data = SoundBOdata)
```

The step function performs an automated selection of the terms to compose the final model considering the backwards selection approach based on stepwise deletion of model terms with high p-values (KUZNETSOVA et al., 2015; ZUUR et al., 2009). The simplification of the random structure of the model is performed based on likelihood ratio test (step 2). Non-significant effects are eliminated and the optimal structure is used to form the simplest plausible model for each attribute according to the principle of parsimony given by the default type I levels ($\alpha = 0.10$ for the random-effects) (KUZNETSOVA et al., 2015).

The step function finds the parsimonious random structure and this structure is considered to test the fixed effects (step 3). The fixed effects are incrementally eliminated following the principle of marginality, that is the effect that are contained in any other effects are retained in the model when the effects that they are contained in are found to be significant according to the specified Type I level ($\alpha = 0.05$ for the fixed-effects) (KUZNETSOVA et al., 2015).

The tables from the automated analysis of the initial mixed model can be extracted by typing the commands:

```
stepBO <-step(M1)
stepBO$rand.table
stepBO$anova.table
```

Table 5 and Table 6 present the results from the Step 2 and Step 3 of the automated analysis for the random and fixed effects respectively, for the attribute 4. The column "Eliminated" in Table 5 and Table 6 gives the order on which the effect was eliminated from the initial model. The effect that have the word "kept" in this column are the ones that form the final model according to the principle of

parsimony given by the default type I levels ($\alpha = 0.10$ for the random effects and $\alpha = 0.05$ for the fixed-effects). Therefore, the practitioner should choose the final model as it contains all the significant effects and the tests for these effects are the most powerful.

It can be seen from the Table 5 that there are six significant random effects that may be considered in our final model. From Table 6 it can be seen that there is a significant interaction between Car, SPL and Track. Then all fixed effects have to enter in our final best model. The final model for the attribute 4 is given by:

$$\begin{aligned} y_{ijklm} =& \mu + \delta_j + \lambda_k + \theta_l + \tau_{jk} + \nu_{jl} + \rho_{kl} + \gamma_{jkl} + \\ & a_i + b_{ij} + c_{ik} + d_{il} + p_{im} + g_{ikl} + \epsilon_{ijklm} \end{aligned} \tag{20}$$

$$\begin{aligned} a_i \sim& N(0, \sigma^2_{assessor}), \\ b_{ij} \sim& N(0, \sigma^2_{assessor \times Car}), \\ c_{ik} \sim& N(0, \sigma^2_{assessor \times Track}), \\ d_{il} \sim& N(0, \sigma^2_{assessor \times SPL}), \\ p_{im} \sim& N(0, \sigma^2_{assessor \times replication}), \\ g_{ikl} \sim& N(0, \sigma^2_{assessor \times Track \times SPL}), \\ \epsilon_{ijkl} \sim& N(0, \sigma^2_{error}). \end{aligned}$$

The reduced model (20), which was selected by using the step function, has quite a complex random structure and a multi-way product structure in the fixed part. It would not possible to fit this model using PanelCheck. Probably other packages or softwares would be able to fit it, but the practitioner does not know in advance which model will be the best in terms of having the best estimates. The step function from lmerTest package finds this model automated.

4.3.4 3-way product analysis using SensMixed

PanelCheck is a good tool for mixed modelling of sensory data in simple situations. The lmerTest was a first step in the way to help sensory practitioner to deal with the challenges of applying mixed modelling for situations that are more complexes. However, it still a challenge for most people to deal with the syntax of

Table 5 Likelihood ratio tests for the random effect and their order of elimination for the automated analysis of SoundBO data.

	χ^2	DF	Eliminated	p-value
RepFixed	0.00	1	1	1.000
Part:Car:SPL:Track	0.00	1	2	1.000
Part:Track:Car	0.69	1	3	0.411
Part:Track:SPL	1.76	1	4	0.185
Part	10.90	1	kept	0.001
Part:Track	19.05	1	kept	< 0.001
Part:Car	15.66	1	kept	< 0.001
Part:SPL	8.86	1	kept	0.003
RepFixed:Part	7.61	1	kept	0.006
Part:Car:SPL	32.80	1	kept	< 0.001

kept means the effect was not eliminated due to it's significance.

Table 6 F-tests for the fixed-effects and their order of elimination of the automated analysis for SoundBO data.

	SQ	MS	DF	F	Eliminated	p-value
Track	0.42	0.11	4	10.54	kept	< 0.001
Car	1.35	0.27	5	26.96	kept	< 0.001
SPL	0.15	0.08	2	7.66	kept	0.004
Track:Car	1.16	0.06	20	5.80	kept	< 0.001
Track:SPL	0.19	0.02	8	2.37	kept	0.016
Car:SPL	1.17	0.12	10	11.68	kept	< 0.001
Track:Car:SPL	0.63	0.02	40	1.59	kept	0.019

kept means the effect was not eliminated due to it's significance.

the R packages. In the way to facilitate even more the mixed modelling for complex situation, such as multi-way product structure, unbalanced data and complex error structure, Kuznetsova et al. (2015) presented the new R package SensMixed. With the same techniques as in lmerTest, the SensMixed is even better in several ways. The new package SensMixed do the mixed modelling faster, analysing all attributes simultaneously and present in the output the appropriate visual tool based on effect size described in Chapter 3. Furthermore, the SensMixed provides an intuitive graphical user interface, which makes it an easy-to-use tool for the sensory practitioners.

The initial linear mixed model is the same specified by model (19), but using the intuitive interface of SensMixed the practitioner just need to click and point to specify the model and analyse all attributes simultaneously. In order to fit a mixed model using the SensMixed package, first type in the R console the following lines:

```
library(SensMixed)
SensMixedUI()
```

Then an intuitive graphical user interface will offer a number of options for the mixed effects model building. To specify the model (19) in SensMixed, the following controls in should be selected:

- **error structure = 2-WAY**: replicate effect and interaction between assessor and replication effects
- **product structure** = 3, main product effects and all possible interactions between them;
- **scaling correction = No**

The sequential chi-squared values from the likelihood ratio tests for the ten random effects for all attributes are presented in Table 7. Figure 5 represents the barplot for the chi-squared values (from the stepwise selection process) followed by level of significance of the effect given by the colour of the bar. The bar size is equal to the chi-squared value from the likelihood ratio tests for each random effect and the colour of the bar indicates the corresponding p-value: Yellow: $0.01 < p$-

value < 0.05, Orange: $0.001 < p$-value < 0.01 and Red: p-value < 0.001. A gray bar means that the effect is non significant.

It can be seen from Figure 5 that the 2-way interactions Track:Part are non-significant for 3 out of 8 attributes. The interactions Car:Part, SPL:Part and 3-way interactions Car:SPL:Part are significant for 7 attributes. It means that the assessors (Part) disagree in scoring the products according to these features for this attributes. We observe that the assessor effects (Part) are significant for 3 out of 8 attributes. The 3-way interactions Track:SPL:Part are significant for 6 out of 8 attributes and Track:Car:Part are non-significant for 4 attributes. The Rep effect is non-significant for all attributes. The interactions between Rep and assessors (Rep:Part) are significant for 7 attributes. The barplot presented in Figure 5 and the Table 7 give information about the significance of the random effects. The barplot is a valuable visual tool that helps to investigate quickly, for instance, whether there is a significant effect according to the colours of the bars. On the other hand, if one is interesting in the chi-square values, the table could be very useful.

After the simplification of the random effects, the optimal random structure found for each attribute is used in the process of estimation of the fixed effects. The mixed model (19) considers the multi-way product structure of the data, accounting for the three main effects Car, SPL and Track. This model gives a better insight into the products structure compared to the model 18 where was considered one product effect with 90 levels. In Table 8 the results of F-test for the three main effects (Car, SPL and Track) and all interactions are presented. To complement the interpretation of the F-test results, the SensMixed provides the delta-tilde plot presented in Chapter 3. As an effect size measurement, the delta-tilde estimate can be compared across any attributes.

Table 8 presents the results of the F-test for the three main effects (Car, SPL and Track) and all interactions. Looking into the multi product structure we can see that the 3-way interaction Car:Track:SPL is significant for all attributes except for the attribute 3. The main effect Car and its interactions with the other effects are significant for all attributes. The main effect SPL is significant for 6 of 8 attributes and the main effect Track is significant for 5 attributes.

Table 7 $\sqrt{\chi^2}$-statistics for LRT for random-effects with significance levels for the SoundBO data.

	Att1	Att2	Att3	Att4	Att5	Att6	Att7	Att8
Part:Track	0.72	19.09^{***}	0.08	19.05^{***}	4.44^{*}	1.57	0.00	2.64
Part:Car	0.26	20.44^{***}	24.65^{***}	15.66^{***}	440.59^{***}	3.92^{*}	14.44^{***}	6.36^{*}
Part:SPL	21.77^{***}	23.40^{***}	42.26^{***}	8.86^{**}	6.36^{*}	3.14	12.79^{***}	6.14^{*}
Part:Track:Car	0.00	5.95^{*}	0.00	0.69	0.40	20.40^{***}	10.28^{**}	21.64^{***}
Part:Track:SPL	73.76^{***}	0.00	26.39^{***}	1.76	6.80^{***}	29.94^{***}	68.80^{***}	7.31^{**}
Part:Car:SPL	264.22^{***}	1.64	9.73^{**}	32.79^{***}	24.52^{***}	7.97^{**}	24.52^{***}	24.33^{***}
Part:Car:SPL:Track	0.00	11.45^{***}	3.95^{*}	0.00	0.00	0.00	0.00	1.16
Part	6.95^{**}	2.67	0.29	10.89^{***}	9.18^{**}	0.00	1.76	0.54
Rep	0.00	0.00	2.92	0.00	0.00	0.15	0.00	0.00
Rep:Part	89.56^{***}	23.83^{***}	0.06	7.61^{**}	32.81^{***}	46.04^{***}	24.88^{***}	21.71^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

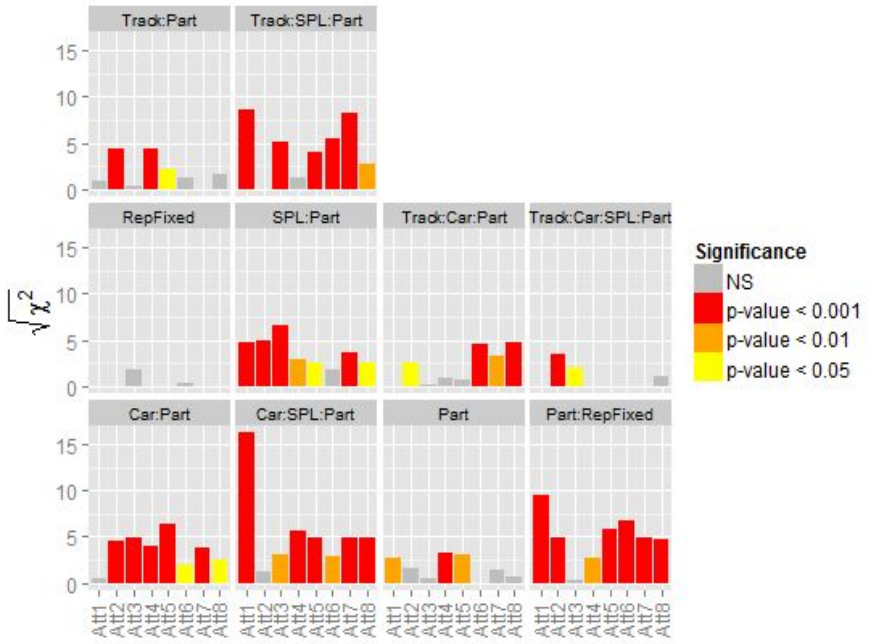

Figure 5 Barplot for $\sqrt{\chi^2}$ for random effects of Sound data.

Table 8 F-test for the fixed effects for SoundBO data

	Car	Track	SPL	Car:Track	Car:SPL	Track:SPL	Car:Track:SPL
Att1	41.56 ***	2.03	20.62***	6.30***	25.91***	2.75**	1.94***
Att2	15.76 ***	1.29	1.85	1.64*	12.60***	0.59	1.58*
Att3	31.71 ***	2.34	3.98*	6.41***	18.87***	1.42	1.37
Att4	26.96 ***	10.54***	7.66**	5.79***	11.68***	2.37*	1.59*
Att5	18.35 ***	6.70***	2.58	4.35***	9.52***	2.73*	1.70**
Att6	19.75***	6.32***	4.27*	5.38***	10.94***	0.85	3.57***
Att7	48.70***	3.31*	12.45***	10.05***	17.48***	2.58*	2.66***
Att8	24.85***	2.95*	23.08***	6.38***	16.70***	1.61	1.86**

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Figure 4 presents the multi-attribute barplots of F-statistics for the fixed effects of the model (18) combined with colour coding of the significance results from the output of PanelCheck. In this way, the F-statistic was used as a kind of effect size measure. This can be a good approach, especially within Panelcheck, where the multi-attribute barplot of the overall product differences are used only for single-factor product effects and with the same choice of F-test denominator across all the attributes of a plot. However, the F-statistic itself is not generally the best measure of effect size as it depends on the number of observations for each product. Furthermore, the various ANOVA mixed models, that we often use for such analysis also complicates the relative effect size handling as generally in mixed models, different effects may have different noise structures, that is, different factors may be tested using different F-test denominators. In this way, SensMixed provide a better multi-attribute plot for sensory and consumer data based on an effect size expressed in terms of relative pairwise comparisons, so-called delta-tilde. It has been shown in Chapter 3 that the effect size used here has a close link with the Thurstonian d-prime, and as such is a generic measure that can be interpreted and compared across any attribute and situations. This visual tool complements the interpretation of the F-test results when product are found significant.

In Figure 6, the delta-tilde estimates for the three main effects (Car, SPL and Track) and all interactions combining with the colour coding of the significance results from the ANOVA output are presented. As we have shown in Chapter 3, the delta-tilde estimates are effects size measures, and then the bars of the delta-tilde plot can be compared. From the Figure 6, it can be seen that the heights of the bars corresponding to the Track and SPL effects and the interaction between them are lower than those pertaining to the Car effect, which means that, the size of the Track and SPL effects are lower than the size of the Car effect. It means that Car effect has a higher impact than Track and SPL on the ability of assessors to discriminate between the products. However, the effect of SPL for the attribute 1 is highly significant and the size of its bar is much higher than for the other attributes, so there is a high impact of the SPL feature on the ability to discriminate between the products for this attribute.

To complement the results of the mixed model analysis of variance, when product effects are found significant, *post hoc* test are performed, also called multiple comparison tests. Figures 7, 8 and 9 present the barplots for multiple comparisons tests together with the 95% confidence intervals for attribute 4 to the main effects Car, Track and SPL respectively.

The six levels of the Car effect are compared 2-by-2 and presented in Figure 7. It is possible to identify which levels of Car are different. It can be seen that all "products" with different levels for Car are different except for levels 1-2, 1-5 and 3-4. The five levels of Track effect are compared 2-by-2 and presented in Figure 8. It can be seen that all "products" with level 4 differ from all other levels of Track (1, 2, 3 and 5). The "products" with level 1 and 3 for Track effect also differ. And the three levels of SPL are compared 2-by-2 and presented in Figure 9. It can be seen that the "products" with level 1 differ from the other levels (2 and 3) of SPL.

The parsimonious model provided by SensMixed for this example has quite a complex random structure and a multi-product structure in the fixed part. Considering the simple model (18) with one-way product structure (combination of Car, Track and SPL, which would result in one fixed product effect CLIP) would not provide this valuable insight into the data. Besides that, the delta-tilde plot for the fixed effects makes the bar heights better comparable for factors with differences in number of levels.

4.3.5 Advanced mixed modelling for sensory data using SensMixed

We can improve even more the analysis of SoundBO data accounting for the scaling difference by using the Mixed Assessor Model (MAM). According with Brockhoff, Schlich and Skovgaard (2015) whenever the scaling is significant it is advisable to correct for it in order to obtain more powerful tests for products. In Brockhoff, Schlich and Skovgaard (2015) the MAM was presented in a simple 2-way structure (cf. 2.2.2). SensMixed package provides an option to correct for scaling, considering more complex structures such as 3-way, where replicate/session effect forms also part of the model as well as multi-way product structures.

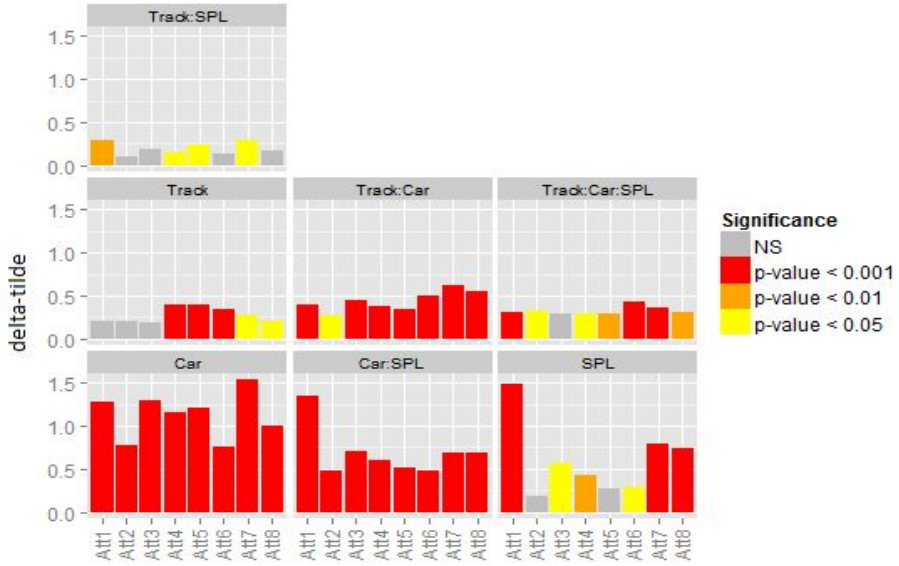

Figure 6 Barplot for delta-tilde estimate of fixed effects of Sound data.

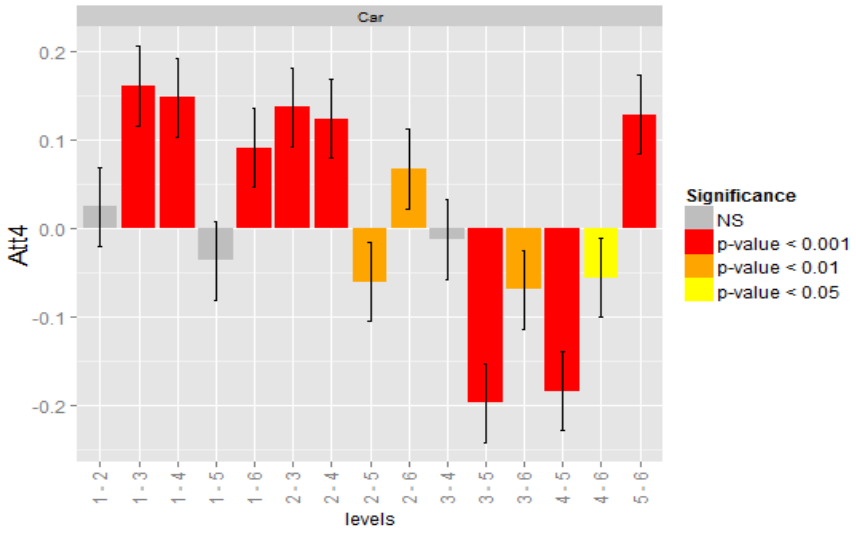

Figure 7 Barplot for differences of least squares means together with the 95% confidence intervals for Car effect of Att 4.

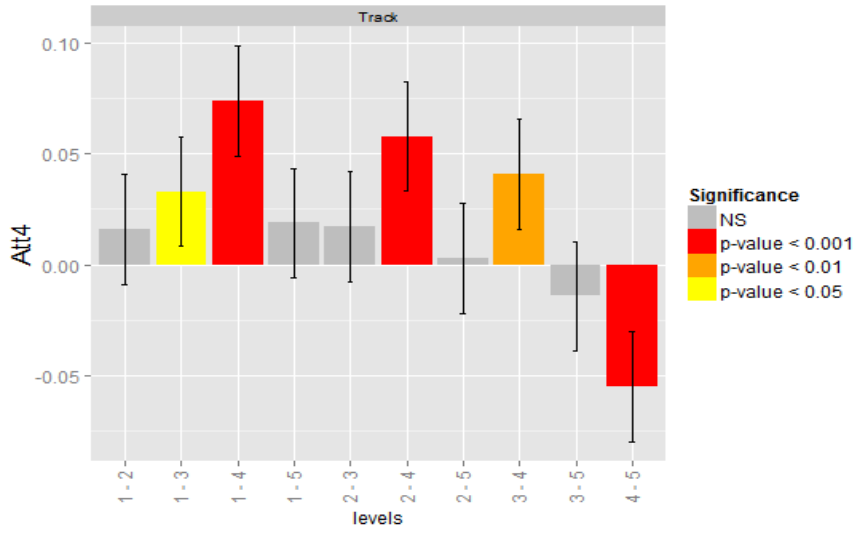

Figure 8 Barplot for differences of least squares means together with the 95% confidence intervals for track effect of Att 4.

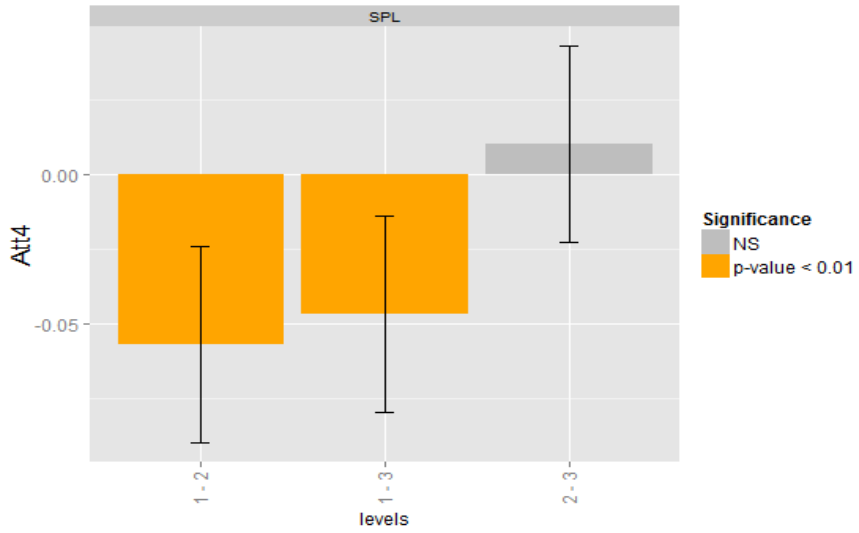

Figure 9 Barplot for differences of least squares means together with the 95% confidence intervals for SPL effect of Att 4.

The mixed assessor model (MAM) that takes individual scaling differences into account for SoundBO can be specified in the following form:

$$\begin{aligned} y_{ijklm} =& \mu + a_i + \delta_j + \lambda_k + \theta_l + \tau_{jk} + \nu_{jl} + \rho_{kl} + \gamma_{jkl} + \\ & \beta_i x_j + d_{ij} + r_m + s_{im} + t_{jm} + \epsilon_{ijklm} \end{aligned} \tag{21}$$

$$\begin{aligned} a_i \sim& N(0, \sigma^2_{assessor}), \\ r_m \sim& N(0, \sigma^2_{replication}), \\ d_{ij} \sim& N(0, \sigma^2_{disagreement}), \\ s_{im} \sim& N(0, \sigma^2_{assessor \times replication}), \\ t_{jm} \sim& N(0, \sigma^2_{product \times replication}), \\ \epsilon_{ijkl} \sim& N(0, \sigma^2_{error}). \end{aligned}$$

where y_{ijklm} corresponds to the attribute in study. The main fixed effects Car, Track and SPL are represented by δ_j, λ_k, θ_l accordingly; the two-way interactions between the fixed effect are represented by τ_{jk}, ν_{jl} and ρ_{kl}; and the three-way interaction is represented by γ_{jkl}. The random part of the model is compounded of the main effect Assessor represented by a_i; the main effect replicate represented by r_m; the interactions between replicate and Assessor represented by s_{im}; plus the interactions between replicate and "product" represented by t_{jm}. We may observe that the random part does not account for the multi-way product structure. The $x_j = \overline{y}_{.j.} - \overline{y}_{...}$ are the centered product averages inserted as a covariate, and β_i is the individual (scaling) slope (with $\sum_{i=1}^{I} \beta_i = 0$), the d_{ij} is the random interaction term, that captures the disagreements between the assessors (BROCKHOFF; SCHLICH; SKOVGAARD, 2015).

To specify the model (4.3.5) in SensMixed, the following controls in should be selected:

- **error structure = 3-WAY**: assessor and replication effect and interaction between them and interaction between them and product effects
- **product structure** = 3, main product effects and all possible interactions between them;

- **scaling correction** = YES
- **One-way product MAM** = YES

Figure 10 represents the barplot for the chi-squared values from the likelihood ratio tests for the random effects of the model (21), followed by level of significance given by the colour of the bar. The colour of the bars (yellow, orange, red) indicate that there is a significant effect and a gray bar means that there is not. The bar size is equal to the chi-squared value from the likelihood ratio tests for each random effect and the colour of the bar indicates the corresponding p-value: Yellow: $0.01 < p$-value < 0.05, Orange: $0.001 < p$-value < 0.01 and Red: p-value < 0.001. From Figure 10 it can be seen that the repetition effect and its interaction with the "product" effect are non-significant for all attributes. However, there is a significant interaction between assessors (Part) and replication for 7 attributes. We can observe as well that the interaction between assessor and "product' (Product:Part) is significant for all attributes. Since here the mixed assessor models are considered, the Product:Part effect means the real disagreement between participants in scoring the products.

Figure 11 shows the $\sqrt{F}$-plot for the scaling effects. From the Figure 11 it is clear that the scaling effect is significant for all attributes, so the participants use the scale differently. Since the MAM is considered, the scaling effect is corrected for. The MAM removes the scaling effect from interaction between assessor and "product" (Product:Part). The consequence for the test of product differences is that the disagreement mean square becomes the one to use in the denominator, improving hypothesis tests for product effects.

Figure 12 represents the delta-tilde estimates for the fixed effect. We can observe that for the MAM (model 21)where the scaling is accounted for, the main effects of Car, SPL and Track are significant for all attributes. The interaction between Car and the other effects are highly significant for all attributes. The heights of the bars corresponding to the SPL and Track effects and the interaction between them are lower than those pertaining to the Car effect, which means that the size of the SPL and Track effects are lower than the size of the Car effect. However, for the attribute 1 the effect of SPL is highly significant and the size of the bar is much higher than for the other attributes, so there is a high impact

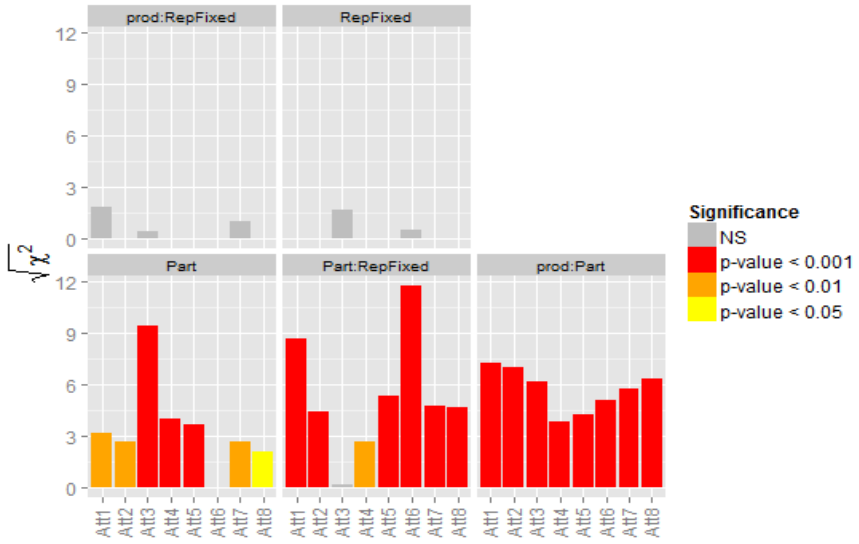

Figure 10 Barplot for $\sqrt{\chi^2}$ of likelihood ratio test for random-effects for SoundBO data

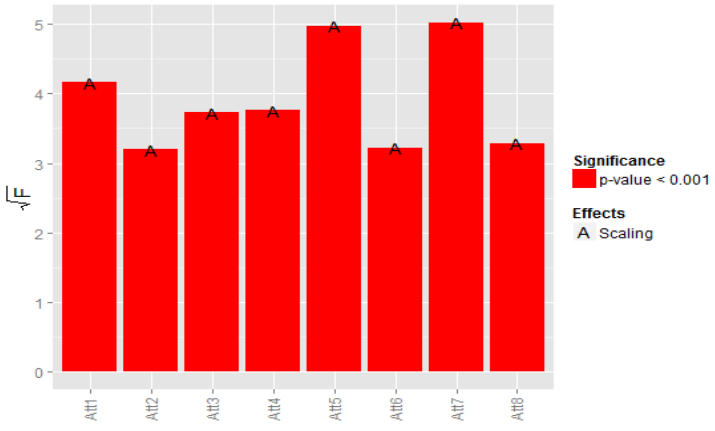

Figure 11 Barplot for $\sqrt{F}$ statistics for fixed-effects scaling for SoundBO data

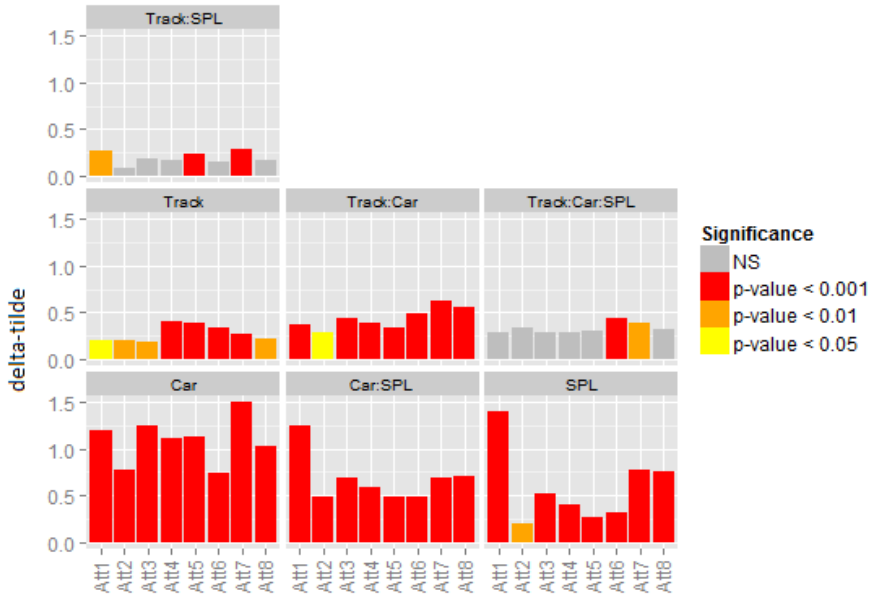

Figure 12 Barplot for delta-tilde estimates for fixed-effects for SoundBO data

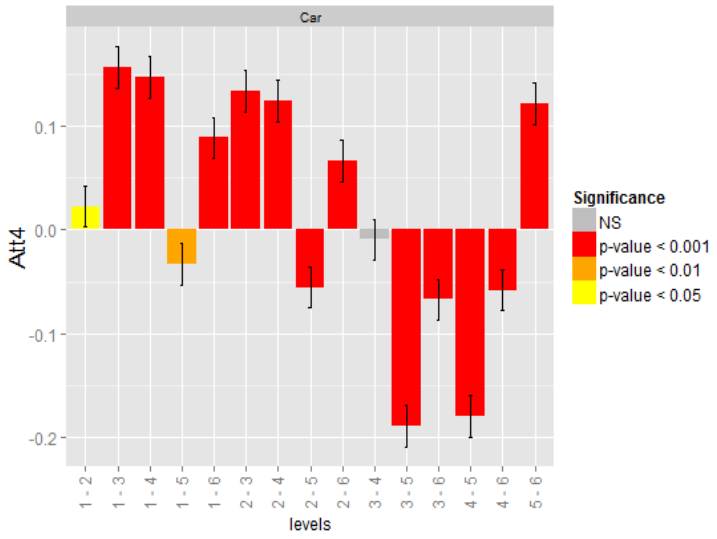

Figure 13 Barplot for differences of least squares means together with the 95% confidence intervals for Car effect of Att 4 for MAM.

of the SPL feature on the ability to discriminate between the products for this attribute. The 3-way interaction between Car, SPL and Track is significant for only 2 attributes.

According to Brockhoff, Schlich and Skovgaard (2015) the MAM produces valid and improved hypothesis tests for as well overall product differences as *post hoc* product difference testing. For instance, the barplots for multiple comparisons tests together with the 95% confidence intervals for attribute 4 to the main effects Car is presented in Figure 13. The six levels of the Car effect are compared 2-by-2 and presented in Figure 13. It is possible to observe that in the MAM all "products" with different levels for Car are different except for levels 3-4.

4.4 Discussion

In this Chapter the PanelCheck software and the two R packages lmerTest and SensMixed were presented as important tools to facilitate the mixed modelling analysis of sensory and consumer data.

PanelCheck can be seen as a very useful open source software to analyse sensory data on simple situations, since the scope of the mixed modelling in PanelCheck is limited in several ways. As we shown in the example, the PanelCheck cannot handle with the multi-way product structure. As an alternative to analyse more complex situations the lmerTest package developed by (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a) was presented.

In lmerTest, automated model selection procedures are available to facilitate the access to the proper mixed modelling for challenging structured situations. We also showed that the lmerTest package provides different kinds of tests on lmer objects (from lme4 package). The example have shown that by using lmerTest the user can perform tests for random and fixed effects for linear mixed effect models. The tests comprise type III and type I F-tests for fixed effects, likelihood ratio tests for random effects and corresponding plots. The package also provides the p-values calculated from F-statistics with Satterthwaite (default) or Kenward-Roger approximations for denominator degrees of freedom . As we shown in the example, the model considering the multi-way product structure has improved the

analysis to achieve more insight of the data.

The lmerTest considerably facilitates the mixed modelling, which are considered the most appropriate choice for a great range of consumer and sensory studies. However, it remains a challenge for most sensory practitioner to deal with the syntaxes of the R packages. In that sense, the new package SensMixed, developed by Kuznetsova, Brockhoff and Christensen (2014b), is even better, since it provides an intuitive graphical user-interface together with all generality of the lmerTest. That makes the SensMixed an easy-to-use tool for the sensory practitioners. As we have shown in the example, the SensMixed provides table and plots to interpret the results, including the barplot for delta-tilde estimates, the appropriate visual tool based on effect size presented in Chapter 3. Furthermore, the SensMixed has the option to specify the Mixed Assessor Model (BROCKHOFF; SCHLICH; SKOVGAARD, 2015), which does the correction for scaling, making the analysis for product more powerful.

In that way, PanelCheck has shown to be a valuable tool for mixed modelling, but just for simple cases. The lmerTest is more general, since it allows for multi-way product structures, incomplete data and complex errors structures. Beyond that it provides the type III ANOVA output with degrees of freedom corrected F-tests for fixed-effects, which makes it the lmerTest unique among the open source softwares. In order to facilitate even more the mixed modelling for complex situation, the SensMixed can be used, as it requires no skills in R-programming and provides advanced statistical methods for analysing sensory data, such as multi-way product structure, unbalanced data and complex error structure. Besides that,SensMixed analyse many attributes in a faster way, since the package is programmed to do the calculation in parallel, which means the analysis of all attributes is made simultaneously. Furthermore, the SensMixed allows improving the interpretation of the ANOVA results with the new visual tool based on the delta-tilde estimates. All that makes the SensMixed package a very valuable tool for sensory practitioners.

5 CONCLUSIONS AND FUTURE PERSPECTIVES

This thesis has a contribution to the field of experimental statistics in general; and to complement results of ANOVA F-test for sensory data in particular. The aim of this thesis was to develop a visual tool to supplement the initial overall ANOVA F-testing, based on effect size estimates. Typically the ANOVA F-testing focus only on the p-values to conclude about the significance of an effect. However, the p-values itself is not useful for estimating the magnitude of the effects. The methodology presented in this thesis suggest the use of an effect size measure, so called delta-tilde, as a visual tool to improve the interpretation of F-test results.

The delta-tilde, as an effect size measure, can be seen as a generic measure that can be interpreted and compared across any attribute and situation. Although the delta-tilde plot suggested here cannot substitute a good *post hoc* analysis, they are valuable additional tools for a good and relevant interpretation of the ANOVA results. In addition, the delta-tilde plot can help to move the focus a bit away from purely looking at p-values but rather focusing on the size of the effect. The application of the delta-tilde plot becomes particularly useful in situations with more than a single factor, as it makes available the comparison of the bar heights for factors with differences in number of levels.

For now, a simple transformation on the F-statistics from ANOVA was used to obtain the effect size measures. Working in the development of significance statements and confidence intervals for the effects could be a relevant following step. The approach presented here transforms the bar plots of F-statistics by re-scaling the bar heights and gives the average pairwise delta-tilde between product levels. It has been showed that it has the same interpretation as the real d-prime calculated from Thurstonian approach. The implications of this relation from the sensmometric point of view, would be interesting to investigate further.

A simple and easy to reproduce methodology for analyzing sensory data in complex mixed modelling framework was presented using flexible and graphically oriented statistical softwares. In addition, an implementation of the delta-tilde plot is available in the R package SensMixed (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014b). The methodology presented in this work can be very help-

ful to sensory practitioners interested in applying mixed models to analyse sensory data.

REFERENCES

AMERINE, M. A.; PANGBORN, R. M.; RESSSLER, E. B. **Principles of sensory evaluation of food**. New York: Academic, 1965. 602 p.

AMORIM, I. S. et al. Linear mixed effects modelling for multifactorial sensory and consumer data using the R - packages lmertest and Sensmixed. In: INTERNATIONAL BIOMETRIC CONFERENCE, 27., 2014, Florence. **Proceedings...** Florence: IBS, 2014. p. 93.

BATES, D. et al. **lme4**: linear mixed-effects models using Eigen and S4. R Package Version 1.1-7. 2014. Available from: <http://CRAN.R-project.org/package=lme4>. Access in: 25 Jan. 2015.

BROCKHOFF, P. B. Sensometrics. In: LOVRIC, M. (Ed.). **International encyclopedia of statistical science**. New York: Springer, 2011. part 19, p. 1302-1305.

BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. Thurstonian models for sensory discrimination tests as generalized linear models. **Food Quality and Preference**, Barking, v. 21, n. 3, p. 330–338, Apr. 2010.

BROCKHOFF, P. B.; SCHLICH, P.; SKOVGAARD, I. Taking individual scaling differences into account by analyzing profile data with the mixed assessor model. **Food Quality and Preference**, Barking, v. 39, p. 156–166, Jan. 2015.

CHANG, W. et al. **shiny**: web application framework for R. R Package Version 0.11.1. 2015. Available from: <http://CRAN.R-project.org/package=shiny>. Access in: 30 Apr. 2015.

CHOW, S. L. Précis of statistical significance: rationale, validity and utility. **Behavioral and Brain Sciences**, Cambridge, v. 21, p. 169–239, 1998.

CHRISTENSEN, R. H. B. **ordinal**: regression models for ordinal data. R Package Version 2014.11-14. 2014. Available from: <http:// www.cran.r-project.org/package=ordinal/>. Access in: 31 Jan. 2015.

CHRISTENSEN, R. H. B.; BROCKHOFF, P. B. **sensR**: an R-package for sensory discrimination. R Package Version 1.4-5. 2015. Available from: <http://www.cran.r-project.org/package=sensR/>. Access in: 31 Mar. 2015.

CHRISTENSEN, R. H. B.; CLEAVER, G.; BROCKHOFF, P. B. Statistical and Thurstonian models for the A-not A protocol with and without sureness. **Food Quality and Preference**, Barking, v. 22, p. 542–549, 2011.

COE, R. It's the effect size, stupid: what effect size is and why it is important. **Annual Conference of the British Educational Research Association**, London, v. 1, p. 12–14, Sept. 2002.

COHEN, J. Earth is round ($p < .05$), The. **American Psychologist**, Washington, v. 49, n. 12, p. 997–1003, 1994.

COHEN, J. Power prime, A. **Psychological Bulletin**, Washington, v. 112, n. 1, p. 155–159, 1992.

COHEN, J. Things I have learned (so far). **American Psychologist**, Washington, v. 45, n. 12, p. 1304–1312, 1990.

CUMMING, G.; FINCH, S. Inference by eye: confidence intervals and how to read pictures of data. **American Psychologist**, Washington, v. 60, n. 2, p. 170–180, 2005.

DAHL, T.; TOMIC, O.; NÆS, T. Some new tools for visualising multi-way sensory data. **Food Quality and Preference**, Barking, v. 19, n. 1, p. 103–113, 2008.

DEVANEY, T. A. Statistical significance, effect size, and replication: what do the journals say? **The Journal of Experimental Education**, Washington, v. 69, n. 3, p. 310–320, Apr. 2001.

ENNIS, D. M. The power of sensory discrimination methods. **Journal of Sensory Studies**, Westport, v. 8, n. 4, p. 353–370, Dec. 1993.

ENNIS, D. M. Thurstonian models for intensity ratings. **IFPress**, Rome, v. 2, n. 3, p. 2–3, 1999.

FAN, X. Statistical significance and effect size in education research: two sides of a coin. **The Journal of Educational Research**, Washington, v. 94, n. 5, p. 275–282, 2010.

FISHER, R. A. The mathematics of a lady tasting tea. In: NEWMAN, J. R. (Ed.). **The world of mathematics**. New York: Simon and Schuster, 1956. p. 1512-1521.

GREEN, D. M.; SWETS, J. A. **Signal detection theory and psychophysics**. New York: J. Wiley, 1966. 272 p.

GRISSOM, R. J.; KIM, J. J. **Effect sizes for research**: univariate and multivariate applications. 2nd. ed. New York: Taylor & Francis, 2012. 430 p.

HAUTUS, M. J.; O'MAHONY, M.; LEE, H. S. Decision strategies determined from the shape of the same-different roc curve: what are the effects of incorrect assumptions? **Journal of Sensory Studies**, Westport, v. 23, p. 743–764, 2008.

HOUT, D. van. **Measuring meaningful differences**: sensory testing based decision making in an industrial context: applications of signal detection theory and thurstonian modelling. 2014. 140 p. Thesis (Ph.D in Philosophiae) — University Rotterdam, Rotterdam, 2014.

HUBBARD, R.; LINDSAY, R. M. Why p values are not an useful measure of evidence in statistical significance testing. **Theory and Psychology**, London, v. 18, n. 1, p. 69–88, 2008.

HUSSON, F.; LE, S.; CADORET, M. **SensoMineR**: sensory data analysis with R. R Package Version 1.20. 2014. Available from: <http://CRAN.R-project.org/package=SensoMineR>. Access in: 17 Jan. 2015.

KELLEY, K.; PREACHER, K. J. On effect size. **American Psychological Association**, Washington, v. 17, n. 2, p. 137–152, 2012.

KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. **lmerTest**: tests for random and fixed effects for linear mixed effect models: lmer objects of lme4 package. R Package Version 2.0-11/r63. 2014a. Available from: <http://R-Forge.R-project.org/projects/lmertest/>. Access in: 23 Mar. 2015.

KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. **SensMixed**: mixed effects modelling for sensory and consumer data. R Package Version 2.0-5/r68. 2014b. Available from: <http://R-Forge.R-project.org/projects/lmertest/>. Access in: 10 June 2015.

KUZNETSOVA, A. et al. Analysing sensory data in a mixed effects model framework using the R package SensMixed. **Food Quality and Preference**, Barking, 2015. In press.

KUZNETSOVA, A. et al. Automated mixed ANOVA modeling of sensory and consumer data. **Food Quality and Preference**, Barking, v. 40, p. 32–38, 2015.

LAWLESS, H. Commentary on random vs fixed effects for panelists. **Food Quality and Preference**, Barking, v. 9, n. 3, p. 163–164, May 1998.

LAWLESS, H. T.; HEYMANN, H. **Sensory Evaluation of Food**: principles and practices. 2nd. ed. New York: Chapman & Hall, 2010. 596 p.

LUNDAHL, D.; MACDANIEL, M. The panelist effect: fixed or random? **Journal of Sensory Study**, Westport, v. 3, p. 113–121, 1988.

MACKAY, D. B.; ZINNES, J. L. A probabilistic model for the multidimensional scaling of proximity and preference data. **Marketing Science**, Greenvale, v. 5, n. 4, p. 325–344, 1986.

MEILGAARD, M. C.; CIVILLE, G. V.; CARR, B. T. **Sensory evaluation techniques**. 4th. ed. Florida: CRC, 2006. 464 p.

MOSKOWITZ, H. R. **Product testing and Sensory Evaluation of Foods**. Westport: Food and Nutrition, 1983. 605 p.

MOSKOWITZ, H. W.; SILCHER, M. The applications of conjoint analysis and their possible uses in sensometrics. **Food Quality and Preference**, Barking, v. 17, n. 4, p. 144–165, Mar. 2006.

NÆS, T.; BROCKHOFF, P. B.; TOMIC, O. **Statistics for Sensory and Consumer Science**. Chichester: J. Wiley, 2010. 287 p.

NÆS, T.; LANGSRUD, O. Fixed or random assessors in sensory profile? **Food Quality and Preference**, Barking, v. 9, n. 3, p. 145–152, May 1996.

NOFIMA, N.; ÅS, A. **PanelCheck software**. 2008. Available from: <www.panelcheck.com>. Access in: 10 May 2015.

O'MAHONY, M. Salt taste sensitivity: a signal detection approach. **Perception**, Ottawa, v. 1, n. 4, p. 459–464, 1972.

O'MAHONY, M. **Sensory evaluation of food**: statistical methods and procedures. New York: M. Dekker, 1986. 487 p.

O'MAHONY, M. Short-cut signal detection measures for sensory analysis. **Journal of Food Science**, Chicago, v. 44, p. 302–303, 1979.

O'MAHONY, M. Who told you the triangle test was simple? **Food Quality and Preference**, Barking, v. 6, n. 4, p. 227–238, Jan. 1995.

STEIGER, J. H. Beyond the f test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. **Psychological Methods**, Washington, v. 9, n. 2, p. 164–182, 2004.

STONE, H.; SIDEL, J. L. **Sensory evaluation practices**. Davis: Elsevier Academic, 2004. 377 p.

SUN, S.; PAN, W.; WANG, L. L. A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. **Journal of Educational Psychology**, Arlington, v. 102, n. 4, p. 989–1004, 2010.

THURSTONE, L. L. A law of comparative judgment. **Psychological Review**, Washington, v. 34, n. 4, p. 273–286, 1927.

TOMIC, O. et al. ConsumerCheck: a software for analysis of sensory and consumer data. **Journal of Statistical Software**, Los Angeles, 2015. In press.

WARNOCK, A. R.; SHUMAKER, A. N.; DELWICHE, J. F. Consideration of Thurstonian scaling of ratings data. **Food Quality and Preference**, Barking, v. 17, n. 7/8, p. 556–561, Dec. 2006.

YATES, F. The influence of statistical methods for research workers on the development of the science of statistics. **Journal of the American Statistical Association**, New York, v. 46, n. 253, p. 19–34, 1951.

ZUUR, A. F. et al. **Mixed effects models and extensions in ecology with R**. New York: Springer Science+Business Media, 2009. 574 p.

APPENDIX A

Proof of the relation between $\tilde{\delta}$ and Cohen's f in the balanced one-way ANOVA setting:

From the basic relation between the sum of squared deviations from the mean and the sum of squared pairwise differences, that we state here without proof:

$$\sum_{i=1}^{I}(\mu_{-}\bar{\mu})^2 = \sum_{i_1<i_2}^{I}(\mu_{i_1}-\mu_{i_2})^2/I,$$

it follows that the definition of $\tilde{\delta}$:

$$\tilde{\delta} = \sqrt{\frac{2}{I(I-1)}\sum_{i_1<i_2}^{I}\left(\frac{\mu_{i_1}-\mu_{i_2}}{\sigma}\right)^2},$$

also can be expressed as:

$$\tilde{\delta} = \sqrt{\frac{2}{(I-1)}\frac{\sum_{i=1}^{I}(\mu_i-\bar{\mu})^2}{\sigma^2}}$$

Hence, we have proved that

$$\tilde{\delta} = \sqrt{2}\Psi = \sqrt{2}f.$$

Proof for the more general bias corrected back transformation for balanced interaction effects:

Considering γ_{ij} as the different interaction contributions, the interaction effect is estimated by

$$\hat{\gamma}_{ij} = \bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot},$$

with $K = I \cdot J$ and $n_1 = \cdots = n_K = n$.

$$F \approx \frac{n\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}^2/DF + \sigma^2}{\sigma^2}$$

where DF is the interaction degrees of freedom defined by $(I-1)(J-1)$.

$$= n\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{\gamma_{ij}}{\sigma}\right)^2/DF + 1$$

We use the basic relation between squared paired differences and the sum of square again:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\gamma}_{ij})^2 = \sum_{ij\neq i'j'}^{K}(\hat{\gamma}_{ij}-\hat{\gamma}_{i'j'})^2/(2K)$$

Taking only the lower triangular part of the difference matrix we have

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\gamma}_{ij})^2 = \sum_{ij<i'j'}^{K}(\hat{\gamma}_{ij}-\hat{\gamma}_{i'j'})^2/(K)$$

And then

$$F \approx \frac{n}{2}\frac{DF}{K-1}\sum_{ij<i'j'}^{K}\left(\frac{\gamma_{ij}-\gamma_{i'j'}}{\sigma}\right)^2/\left(K(K-1)/2\right)+1$$

$$= \frac{n}{2}\frac{K-1}{DF}\,(\text{Average squared pairwise dprimes}) + 1$$

And hence, for the interaction we have:

$$\tilde{\delta} = \sqrt{\frac{2}{n}}\sqrt{\frac{DF}{K-1}}\sqrt{F-1}.$$

APPENDIX B

R-code for the analysis of the TVbo data set in SensMixed package

The mixed model for one attribute y_{ijkl} can be specify as:

$$\begin{aligned} y_{ijkl} = \mu + \tau_j + \rho_k + \gamma_{jk} \\ + a_i + b_{ij} + c_{ik} + d_{ijk} + \epsilon_{ijkl} \end{aligned} \tag{22}$$

$$\begin{aligned} a_i &\sim N(0,\sigma^2_{assessor}) \\ b_{ij} &\sim N(0,\sigma^2_{assessor \times TV set}) \\ c_{ik} &\sim N(0,\sigma^2_{assessor \times Picture}) \\ d_{ijk} &\sim N(0,\sigma^2_{assessor \times TV set \times Picture}) \\ \epsilon_{ijkl} &\sim N(0,\sigma^2_{error}) \end{aligned}$$

The fixed part of the model contains a multi-way product structure given by τ_j, ρ_k and γ_{jk} that represents the effect of TVset and Picture and the interaction TVset:Picture respectively. The random part of the model is compounded of the main effect Assessor represented by a_i plus the interactions between Assessor and fixed effects (TVset and Picture and the interaction TVset:Picture) given by b_{ij}, c_{ik} and d_{ijk}.

Using the SensMixed package we can analyse the 15 attributes in TVbo data with a few command lines. First we attach the SensMixed package (version 2.0-7) by typing the following command in the R console:

```
library(SensMixed)
```

The TVbo data set is available in the SensMixed package. To access the TVbo data use the command:

```
data(TVbo)
```

Then we use the function sensmixed to construct the mixed model for all attributes:

```
resTV <- sensmixed(attributes=names(TVbo)[5:ncol(TVbo)],
                       Prod_effects=c("TVset", "Picture"),
                       individual="Assessor",
                       calc_post_hoc = TRUE,
```

```
product_structure=3,
error_structure ="No_Rep",
reduce.random=FALSE,
parallel=FALSE,
data=TVbo)
```

The sensmixed function contains a lot of arguments, here we explain the arguments used above:

- attributes: a vector containing the names of the sensory attributes
- Prod_effects: names of the variables related to the product
- individual: name of the column in the data that represent assessors
- data: data frame (data from sensory studies)
- product_structure: one of the values in $1, 2, 3$.
  - 1: only main effects will enter the initial model.
  - 2: main effects and 2-way interaction.
  - 3: all main effects and all possible interaction.
- error_structure ="No_Rep": assessor effect and all possible interactions between assessor and product effects.

The mixed models for each attribute are constructed using the lme4 package (BATES et al., 2014) and then the step method from the lmerTest (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2014a) is applied to each model. By default the non-significant random effects are eliminated from the model according to the specified by a user Type 1 error (KUZNETSOVA et al., 2015). However to estimate the delta-tilde and compare the bars of the plot, the elimination of the random effects is not required. It can be done by the argument reduce.random=FALSE. By default the computation is done in parallel Kuznetsova et al. (2015). Here we chose parallel=FALSE.

The sensmixed function provides us with the tables of the random and fixed part of the model as well the bar plot presented in the section 3.5. To get the results we simply type the following into R console:

```
resTV
plot(resTV, dprime=TRUE, isRand = FALSE)
```

R-code to obtain the d-prime from Ordinal package

First we attach the packages by the typing the following command in the R console:

```
library(ordinal)
```

Then categorize the subset of TVbo data:

```
TVbo$Cutting_ord1=
    as.integer((cut2((TVbo$Cutting-1)/(max(TVbo$Cutting)-1),
    cuts=(0:10)/10)))
TVbo$Cutting_ord2=factor(TVbo$Cutting_ord1,ordered=TRUE)
```

And finally use the clm function to obtain the d-prime estimate:

```
clm <- clm(Cutting_ord2 ~ TVset,link="probit",
            data = subset(TVbo, TVset != "TV2"))
coef(clm)[-(1:9)]
round(coef(clm2),2)
```