Individual differences in replicated multi-product experiments with Thurstonian mixed models for binary paired comparison data

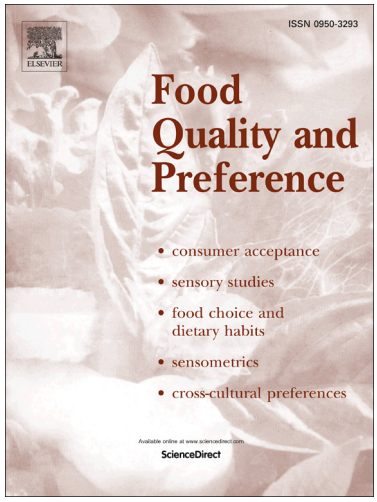
Linander, Christine Borgen; Christensen, Rune Haubo Bojesen; Cleaver, Graham; Brockhoff, Per Bruun

Published in:
Food Quality and Preference

Link to article, DOI:
10.1016/j.foodqual.2019.01.010

Publication date:
2019

Document Version
Peer reviewed version



Individual differences in replicated multi-product experiments with Thurstonian mixed models for binary paired comparison data

Christine Borgen Linander, Rune Haubo Bojesen Christensen, Graham Cleaver, Per Bruun Brockhoff

PII: S0950-3293(18)30777-8
DOI: https://doi.org/10.1016/j.foodqual.2019.01.010
Reference: FQAP 3638

To appear in: *Food Quality and Preference*

Please cite this article as: Linander, C.B., Bojesen Christensen, R.H., Cleaver, G., Brockhoff, P.B., Individual differences in replicated multi-product experiments with Thurstonian mixed models for binary paired comparison data, *Food Quality and Preference* (2019), doi: https://doi.org/10.1016/j.foodqual.2019.01.010

Individual differences in replicated multi-product experiments with Thurstonian mixed models for binary paired comparison data

Christine Borgen Linander[a,b,*], Rune Haubo Bojesen Christensen[a,c], Graham Cleaver[d], Per Bruun Brockhoff[a]

[a] *DTU Compute, Section of Statistics and Data Analysis, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark*
[b] *Clinical Research Center, Copenhagen University Hospital, Hvidovre, Kettegård Alle 30, DK-2650 Hvidovre, Denmark*
[c] *Christensen Statistics, Bringetoften 7, DK-3500 Værløse, Denmark*
[d] *Unilever Research and Development, Port Sunlight, Wirral, UK, CH63 3JW (retired)*

Abstract

Often sensory discrimination tests are performed with replications for the assessors. In this paper, we suggest a new way of analyzing data from a discrimination study. The model suggested in this paper is a Thurstonian mixed model, in which the variation from the assessors is modelled as a random effect in a generalized linear mixed model. The setting is a multi-product discrimination study with a binary paired comparison. This model makes it possible to embed the analyses of products into one analysis rather than having to do an analysis for each product separately. In addition, it is possible to embed the model into the Thurstonian framework obtaining d-prime interpretations of the estimates. Furthermore, it is possible to extract information about the assessors, even across the products. More specifically, assessor specific d-prime estimates are obtained providing a way to get information about the panel. These estimates are interesting because they make it possible to investigate if the assessors are assessing in a specific way.

Keywords: Thurstonian modelling, binary paired comparison, assessor information, multi-product setting, Generalized Linear Mixed Model

1. Introduction

It is a recurrent scenario that discrimination tests are conducted with replications for the assessors (Ennis (2012)). Thus, it is important to handle the possible differences between the assessors correctly. Suggestions in the literature are e.g. the so-called beta-binomial models as well as corrected beta-binomial models. In this paper, we suggest a new way of modelling the potential assessor differences.

It has been described in the literature how Thurstonian modelling is the preferred approach to quantify the difference between products (e.g. Ennis (1993), Ennis & Jesionka (2011), Næs et al. (2010)). In this paper, we follow this recommendation, thus we will consider the analyses on the d-prime scale. Hence, we will be considering Thurstonian models.

This work is part of an overall objective of aligning Thurstonian d' analysis with the modern world of statistical modelling. Brockhoff & Christensen (2010) show how a Thurstonian model for sensory discrimination tests can be seen as a Generalized Linear Model (GLM). The way we suggest to handle the possible assessor differences is by adding assessor as a random effect to a GLM. This results in a Generalized Linear Mixed Model (GLMM), which is a way to analyze categorical data like binomial data. Categorical data analysis is a common well-known framework, which is used in many applications. The book by Agresti (2013) gives a thorough description of categorical data analysis.

The setting that is considered is a multi-product setting giving the possibility to investigate for possible assessor-by-product interactions. In discrimination testing test protocols exist where there is no correct answer. The test protocol that is considered in this paper is the binary paired comparison. This allows for the d-prime values to be positive as

*Corresponding author. E-mail address: christine.borgen.linander@regionh.dk (C. B. Linander).

well as negative. In Section 2.2 we will go into details about the Thurstonian model for this setup.
We believe that adding this level of details to the models give us valuable insights about data that would have been undetected otherwise. Not only do we get the d-prime interpretation of our parameters, in addition, we gain information about the assessors. Moreover, it is possible to embed the analysis of the products into one analysis instead of having to do an analysis for each product separately. Furthermore, the replications of the assessors are handled correctly when testing for a significant effect of products.
We consider figures of the assessor specific d-prime values, giving an opportunity to get insights about the assessors, which is only possible due to the level of details in the model. From these figures, it will be possible to gain knowledge about the panel. Additionally, these figures make it possible to realize whether some assessors are having a pattern in how they are assessing. Furthermore, since no correct answer exists it will be possible to detect if the panel is in agreement about which sample had the strongest intensity.
In the remainder of this section a discrimination study is described. In Section 2 we define the methodology we suggest. We will throughout Section 2 illustrate the methodology by using the study described in section 1.1. At the end of this paper we have a discussion in Section 3.

1.1. The discrimination study

In this section an existing discrimination study is explained. We use this as an ongoing example throughout this paper to illustrate the methodology we introduce in Section 2.
The overall aim of this study was to find a new product that has some of the same characteristics as an existing product. In this study, the assessors were comparing different test products to a control product. A sample of a test product as well as a sample of the control product were applied to an assessor's own skin. The assessor had to choose the sample with the strongest intensity of the attribute in question.
The organization of evaluations of the test products is illustrated in Figure 1. In one day, assessors evaluated two test products, where each assessor evaluated each test product twice in two different sessions.
In total eight test products (denoted by `A`, `B`, `C`, `D`, `E`, `F`, `G` and `H`) were compared to the control product.
The assessors that participated in the assessments of the test products were the same from day to day. Some assessors were not able to participate in the assessments for some days. If an assessor assessed the test products within a day, the assessor participated in both sessions carried out that day. For an assessor to be included in the analysis, the assessor had to participate in at least 50% of the assessments. 25 assessors (denoted by `1`, ..., `25`) made enough assessments to be included in the analysis (two assessors did not make enough assessments).

[Figure 1 about here.]

The assessors assessed multiple attributes and their characteristics are listed in Table 1. The assessors evaluated five different attributes all of which were evaluated immediately after application of the samples. In addition, three of these attributes were re-evaluated after five minutes. Thus, in total eight attributes were assessed by the assessors.

[Table 1 about here.]

2. Methodology

In this section we explain the methodology as well as applying this methodology to the data described in Section 1.1. When analyzing such data a model is fitted for each attribute at a time, thus results are obtained for each attribute separately.

2.1. Explorative investigation of data

A way to gain information about the data obtained from a sensory discrimination study, is to examine proportions. In this section an explorative investigation of the data from Section 1.1 is given.
One aim of analyzing the data is to gain knowledge about which (if any) of the test products that have the characteristics that are desired for this type of product. To gain information about which test product that has the most interesting

sensory characteristics we can look at proportions. The proportions, the number of times the test product was chosen as having higher sensory intensity than the control, are aggregated over assessors as well as sessions. These proportions (in percentages) are given in Table 2.

[Table 2 about here.]

An important sensory characteristic is that the test product should be at least as silky as the control product. When a test product was chosen more often than the control, the proportion is larger than 50%. Thus, for a test product to be silkier than the control the proportion must exceed 50%. Since test product `D` is the only product with proportions, initially and after five minutes, which exceed 50% it is the most promising test product with respect to silkiness being silkier than the control.

2.2. d-prime values for test products

When considering the proportions from the previous section an overview of data is given. However, it can be rather difficult to comprehend how similar (or different) the products are. Thus, the proportions are transformed into d-prime values for a better comparison of the products. We will in this section find the d-prime values for the test products to express the sensory difference between the test products and the control for the eight attributes.
These d-prime values are found by transforming the proportion of times the test product was chosen for each attribute via the inverse of the so-called psychometric function.
To develop the Thurstonian model for our setting let C and T denote the distribution of the sensory intensity for the control product and a test product respectively. We assume that C and T are independent and that:

$$C \sim N(\mu_c, \sigma^2) \qquad \text{and} \qquad T \sim N(\mu_t, \sigma^2).$$

The underlying Thurstonian relative sensory difference δ is the difference in means scaled by the common standard deviation:

$$\delta = \frac{\mu_t - \mu_c}{\sigma}.$$

An advantage of using δ; the measure for sensory differences is that δ does not depend on the discrimination test protocol, see e.g. Ennis (1993)
The psychometric function f_{psy} can for this setting be defined as the probability that the test product is chosen which is the probability of the test product having a larger sensory intensity than the control:

$$f_{psy}(\delta) = P(T > C) = \Phi\left(\frac{\delta}{\sqrt{2}}\right) = p \tag{1}$$

where Φ is the cumulative distribution function for the standard normal distribution and p is the probability that the test product is chosen over the control product. The derivation of the psychometric function in (1) is the same as for the 2-AFC test (Brockhoff & Christensen, 2010; Ennis, 1993), except that there is no restriction on the parameter values.
d', the estimate of δ, is the estimated sensory difference between the test product and the control product. d-prime values can be computed using the inverse of the psychometric function:

$$f_{psy}^{-1}(p) = \Phi^{-1}(p)\sqrt{2} = d' \tag{2}$$

A d-prime value for each comparison of a test product to the control is obtained.
When $p = 0.5$, corresponding to a d-prime value of 0, the assessors chose the test product half the time. Thus, there is no perceivable difference between the test product and the control product. When $p > 0.5$ the d-prime value is positive and the psychometric function is the same as for the 2-AFC protocol. Additionally, for all d-prime values the setting corresponds to the paired comparison protocol, which in some situations also is the paired preference (Christensen et al. (2012)). A positive or negative d-prime value corresponds to the test product having the strongest or weakest intensity of the attribute in question.

The d-prime values for the test products, for the eight attributes, are shown in Table 3. As expected from the values of the proportions, `D` is the only test product with a positive d-prime for `Silky` both evaluated initially and after five minutes.

[Table 3 about here.]

2.3. Generalized Linear Models

The d-prime values from Section 2.2 are calculated from the data without other assumptions than those regarding the underlying distributions for the sensory intensities. Another way to gain information about the data is by imposing a model to the probabilities of a test product being chosen. The observations from the binary paired comparison test protocol are binomially distributed:

$$Y_{ijk} \sim \text{binomial}(p_{ij}, 1)$$

where $i = 1, \ldots, l$ represents the test products, $j = 1, \ldots, n_i$ represents the assessors for the ith test product and $k = 1, \ldots, r$ ($r = 2$ and $l = 8$ for the discrimination study used in this paper) represents the sessions carried out on the same day. In addition, we assume that p_{ij}, the probability of the jth assessor choosing the ith test product, is independent of the sessions:

$$p_{ij} = P(Y_{ijk} = 1)$$

It is possible to impose a linear structure of p_{ij} which explains the variables that are affecting these probabilities. One way of defining this linear model structure is by letting the test products be the only variable that affects the probabilities:

$$p_{ij} = f_{psy}(\mu + \alpha_i) \tag{3}$$

where f_{psy} is the psychometric function given in (1). Thus, the psychometric function is describing how the parameters μ and α_i are relating to the probability p_{ij}. According to Brockhoff & Christensen (2010) this way of writing a Thurstonian model is a Generalized Linear Model and we refer the reader to Brockhoff & Christensen (2010) for further details on this matter.

The parameter μ is the average difference between test products and the control product. α_i is the difference for the ith test product to the average product-difference μ. Or put differently, α_i is the magnitude of how much the ith test product is different from the average product-difference. Thus, the relation between the underlying sensory difference δ_i for the ith test product to the control product and the model parameters is:

$$\delta_i = \mu + \alpha_i \tag{4}$$

The d-prime value d_i', the estimate of δ_i given in (4), is the estimated sensory difference between the ith test product and the control product. These estimates can be found using standard statistical software fitting Generalized Linear Models with the probit link. The d-prime values obtained from using model (3) are listed in Table 3. These values are also the values obtained by transforming the proportions in Section 2.2. Thus, analyzing data with a GLM gives the same d-prime values as transforming the proportions. An advantage of using the GLM approach is that the statistical software provides additional information to the d-prime estimates e.g. standard errors and p-values. Furthermore, realizing that a GLM is another way to write the transformation of the proportions, makes it possible to consider other ways of defining the linear model structure.

2.4. Generalized Linear Mixed Model as a Thurstonian Mixed Model

It was, in the previous section, established that the d-prime values are obtainable using a generalized linear model. In this section, the linear model structure is extended to include a random effect. For other applications, an extension of a GLM to include a random effect is known as a Generalized Linear Mixed Model (GLMM). In this section the linear model structure is extended by adding the effect of the assessors as a random component. Thus, this section is

considering a Thurstonian Mixed Model with a fixed effect of test products as well as a random effect of the assessors. The linear model structure for this model reads:

$$p_{ij} = f_{psy}(\mu + \alpha_i + b_j) \tag{5}$$

where i, j, μ and α_i are defined as described in Section 2.3. Furthermore, $b_j \sim N(0, \sigma_b^2)$ is the random effect of the jth assessor which are independent for all j. b_j is the difference for the jth assessor to the average product-difference μ on the d'-scale. Thus, the sensory difference, on the d-prime scale, between the test products and the control product for the jth assessor is $\tilde{b}_j = \mu + b_j$.

The d-prime values for different assessors for the same product can be different. A standard reason for individual d-prime values to be different is differences in perceptual variances. However, within this model the differences are occurring because the assessors use different criterions for choosing which sample has the strongest intensity of the attribute in question.

The relation between the product d-prime value δ_i and the model parameters is not affected by the random effect of the assessors. This is because the value of δ_i is for an average assessor, thus b_j equals 0, hence the relation is the same as in equation (4). The size of d_i', the estimate of δ_i, depends on how the linear model structure is defined. The values of d_i' using the model structures defined in (3) and (5) for `Silky` after 5 minutes are shown in Figure 2.

[Figure 2 about here.]

Generally, the estimates are further away from zero when the effect of the assessors is taken into account.

2.5. Extending the Thurstonian Mixed Model

The model from Section 2.4 considers the main effect of products and assessors. In this section, the Thurstonian Mixed Model given in (5) is extended, such that the interaction of the products and assessors is included in the linear model structure of the probabilities. The assessor-by-product interaction is a random effect because assessor is included as a random effect. Thus, this section is considering a Thurstonian Mixed Model with a fixed effect of products as well as random effects of the assessors and the assessor-by-product interaction. The linear model structure for this model reads:

$$p_{ij} = f_{psy}(\mu + \alpha_i + b_j + d_{ij}) = f_{psy}(\eta_{ij}) \tag{6}$$

where $d_{ij} \sim N(0, \sigma_d^2)$ is the random effect of the interaction of the ith test product and the jth assessor, which are independent for all i and j. d_{ij} is the difference for the jth assessor for the ith test product to the average product-difference μ on the d-prime scale.

The relation between the product d-prime value, δ_i, and the model parameters is not affected by the random effect of the assessors nor the assessor-by-product interaction. This is because the value of δ_i is for an average assessor, thus b_j and d_{ij} are 0 and the relation remains that $\delta_i = \mu + \alpha_i$.

The model defined by (6) relates to other well-known models in the sensory field. The structure of η_{ij} in (6) resembles the usual 2-way mixed structure for sensory profile data. The usual 2-way analysis of sensory profile data can be done in Panelcheck. If we were to consider a setting with only one test product and multiple observations for each assessor, this corresponds to the usual replicated difference test, which can be modelled by e.g. beta-binomial models.

2.6. Simplification of a Thurstonian Mixed Model

It is of interest to investigate the possibility to describe the data with a simpler model. It will become easier to interpret the results in situations with a simpler model e.g. models with a non-significant assessor-by-product interaction. Thus, it is important to consider the tests of the variables that are included in the linear predictor. This section describes how to investigate whether the linear model structure in (6) can be simplified.

The first test that is considered is the test of the assessor-by-product interaction. Both assessor and product effects are nested within the assessor-by-product interaction, thus it is important to consider the test of the interaction before testing for assessor and product effects.

The interpretation of the assessor-by-product interaction is that the differences between the assessors depend on the

products. Therefore, when testing for a significant assessor-by-product interaction it is investigated whether the assessor differences vary with the products. Since the assessor-by-product interaction is a random effect the hypotheses are statements about the variance parameter. For the test of a significant assessor-by-product interaction, the null hypothesis is that the variance equals zero, while the alternative hypothesis is given as the variance being larger than zero:

$$H_0 : \sigma_d^2 = 0 \quad H_1 : \sigma_d^2 > 0 \tag{7}$$

The alternative hypothesis is one-sided since the variance is non-negative; see Christensen & Brockhoff (2013) for details. The distribution of the test statistic is the Chi-squared distribution with 1 degree of freedom.

[Figure 3 about here.]

The likelihood ratio test statistics for the test of a significant assessor-by-product interaction are shown in Figure 3. The eight attributes have non-significant assessor-by-product interactions. Thus, there is no evidence that the differences between assessors depend on the test products.

The model that is used for testing the main effects of assessors and test products is the model without the assessor-by-product interaction. This model is given in (5). When the assessor-by-product interaction is significant, the understanding of the model becomes more difficult. It is a scope of future research how to define and interpret the test of the main effects of products as well as assessors in the case of a significant assessor-by-product interaction.

The hypothesis test of a significant effect of test products investigates whether the difference between the control and the test products is the same for all the test products. The likelihood ratio test statistic becomes $-2\log(Q) = 2\ell_{H_1} - 2\ell_{H_0} \sim \chi^2(l-1)$ (Pawitan (2001)), where ℓ_{H_0} and ℓ_{H_1} are the log likelihood functions under the null and alternative hypothesis respectively. Furthermore for the data used as an ongoing example in this paper $l - 1 = 7$. The model under the alternative hypothesis is given by (5) allowing for the test products to have different sensory characteristics for that attribute. Furthermore, the model under the null hypothesis is stating that the test products are perceived to be similar compared to the control:

$$p_{ij} = f_{psy}(\mu + b_j)$$

The likelihood ratio test statistics for the test of a significant product main effect are shown in Figure 3. For all attributes, the product main effect is significant, meaning that the test products are perceived differently compared to the control for all the attributes.

Currently assessor replication is often ignored in the analysis of these types of studies, e.g. due to limitations of available software. In such analyses the model reads:

$$p_{ij} = f_{psy}(\mu + \alpha_i) \tag{8}$$

where μ and α_i are defined as previously described. The likelihood ratio test of the product main effect is equivalent to the test for the model including assessor. Thus, the model under the null hypothesis reads:

$$p_{ij} = f_{psy}(\mu)$$

The values of the likelihood ratio test statistic, as well as the values for the test with assessor included in the model, are shown in Figure 4. The value of the likelihood ratio test statistic is generally higher for the test when assessor is included in the model. For some attributes, the difference is small, whereas the difference for other attributes is rather large. The size of the likelihood ratio statistics is just as important as the difference between them, regarding the impact of which model is used. `Silky` (0 minutes) and `Greasy` (0 minutes) approximately have the same size of the difference (approximately 8 and 9 respectively). For `Silky` (0 minutes) the difference is unimportant because both values are large. However, the difference for `Greasy` (0 minutes) is important because both values are small. For the 0.01 level the conclusion, for `Greasy` (0 minutes), depends on which model is used; when ignoring the assessor replicates (model (8)) the null hypothesis is not rejected, whereas inclusion of assessors (model (5)) results in a rejection of the null hypothesis. It is a scope of future research to investigate how much the test of product main effect is affected by ignoring the assessor replicates.

[Figure 4 about here.]

The hypothesis test of a significant assessor main effect is considering whether the assessors perceive the test products differently. Thus, the null hypothesis is assuming that the assessors perceive the products similarly, whereas the alternative hypothesis allows for differences between the assessors. The hypothesis test of a significant assessor main effect is equivalent to the hypothesis test of a significant assessor-by-product interaction, with σ_d^2 being replaced by σ_b^2 in (7). The likelihood ratio test statistics for the test of a significant assessor main effect are shown in Figure 3. The attributes `Thickness` and `Absorption` have non-significant assessor main effects. Hence, there is not enough evidence to claim a significant effect of the assessors for these two attributes. Thus, the assessors perceive the test products similarly for `Thickness` and `Absorption`. For the remaining six attributes, the assessor main effect is strongly significant. Therefore, the assessors perceive the test products compared to the control differently for these attributes.

2.7. Product specific d-prime values

It is of interest to find the product specific d-prime values because this will make it possible to compare the sensory characteristics of the different products. The product specific d-primes are estimated from the model without the assessor-by-product interaction. Thus, the Thurstonian mixed model in (5) is used when finding the product specific d-prime values. Therefore, the estimate, on the d-prime scale, for the ith product reads:

$$d_i' = \hat{\mu} + \hat{\alpha}_i$$

where $\hat{\mu}$ and $\hat{\alpha}_i$ are the estimates of μ and α_i. The estimates of μ and α_i are obtainable from the output in the statistical software.

When the assessor-by-product interaction is significant, the interpretation of the product specific d-prime values become more difficult. In the situation with a significant assessor-by-product interaction one must be cautious when interpreting the product specific d-prime values, because these estimates do not contain all information about the products. It is a scope of future research to investigate the interpretation of the product specific d-prime values when the assessor-by-product interaction is significant.

Confidence intervals for the d-prime values can be found using the Wald-based approach. The 95% Wald-based confidence interval for d_i reads:

$$d_i' \pm z_{97.5}\, se(d_i') \tag{9}$$

where $z_{97.5}$ is the 97.5% quantile for the standard normal distribution. Furthermore, $se(d_i)$ is the standard error of d_i. The standard errors are obtained from the output in the statistical software used when analyzing data with a generalized linear mixed model.

[Figure 5 about here.]

The product specific d-prime estimates as well as the 95% confidence intervals for `Sticky`, `Greasy` (0 minutes) and `Silky` (5 minutes) are shown in Figure 5. Test products `A`, `G` and `H` are more sticky than the control product, whereas the remaining test products are less sticky. The test products furthest to the left (`C`, `D` and `E`) are the most promising test products with respect to stickiness, since the desired characteristic is to be less sticky than the control. All the test products are perceived to be less greasy than the control product, since the d-prime values for `Greasy` are negative. All test products are good candidates with respect to greasiness, since a desired characteristic for the new product is not to be greasier than the control product. The only test product that is perceived to be more silky after 5 minutes than the control product, is test product `D`. The d-prime values for test products `C` and `F` are close to 0, which indicates that these are among the most silky test products after 5 minutes. All in all when considering the results for the attributes `Sticky`, `Greasy` (0 minutes) and `Silky` (5 minutes) the most promising test product is test product `D`.

2.8. Assessor specific d-prime values

It is of interest to find the assessor specific d-prime values because these values make it possible to get insights about the assessors. As for the product specific d-prime values the interpretation of the assessor specific d-prime values is more difficult when the assessor-by-product interaction is significant. Thus, d-prime values for the assessors will be calculated using model (5). The average sensory difference between the test products and the control product for the jth assessor is on the d-prime scale:

$$\tilde{b}_j = \mu + b_j \tag{10}$$

The estimate of $\tilde{b}_j$ in (10) is obtained from the output in the statistical software used when analyzing data with a generalized linear mixed model.

For a balanced design the assessor with the smallest value has been choosing the control most often of all the assessors, whereas the assessor with the highest value has been choosing a test product most often. Assessors with a value of 0 have been choosing the control and a test product half of the times each. The assessors with larger values than the consensus (μ) have on average chosen a test product more often than the average. The assessors with smaller values than the consensus have on average chosen the control more often than the average.

The assessor specific d-prime values make it possible to look for overall tendencies of the assessors. It is possible to investigate if some assessors have a tendency of choosing either the control or a test product. Furthermore, it can be detected if an assessor chooses the control and test products half the time each. However, even though two assessors have similar values their patterns for the different test products might be different. For more detailed information other assessor specific d-prime values could be developed.

The d-prime estimates for the assessors, $\tilde{b}_j$, is for `Silky` (0 minutes) shown in Figure 6.

[Figure 6 about here.]

The assessor specific d-prime estimates, $\tilde{b}_j$, are negative for `Silky` evaluated after 0 minutes. Thus, the assessors are in agreement that the control, on average, is silkier than the test products. The d-prime values for the assessors furthest to the left, assessors `9`, `16` and `19`, are close to -2.5, which is rather far away from 0. This implies that these assessors have chosen the control much more than the test products. In addition, these assessors are the assessors with the smallest proportions of times the test products were chosen. There is a group of assessors, from `1` to `21` looking at the y-axis, whose estimates are close to -2. These assessors have larger proportions, of times the test products were chosen, than the group furthest to the left. The assessors from these two groups, the assessors from `9` to `21` looking at the y-axis, have d-prime estimates less than the consensus, the estimate of μ.

Assessor `4` is the only assessor that has a d-prime value equal to the consensus.

There is a group of assessors, from `20` to `7` looking at the y-axis, that have larger d-prime estimates than the consensus. Assessor `7` is the assessor with the d-prime value closest to 0. Thus, assessor `7` is the assessor with the largest proportion, of times the test products were chosen, of the assessors.

The d-prime values for the assessors, $\tilde{b}_j$, is for `Sticky` (0 minutes) shown in Figure 7.

[Figure 7 about here.]

The assessor specific d-prime values for `Sticky` (0 minutes) are negative as well as positive, with the majority being negative. Therefore, some assessors have chosen the test products more often than the control, however the majority of the assessors have chosen the control more often than the test products.

Assessors `1`, `6` and `9` are the assessors with the smallest proportion of times the test products were chosen. Assessor `24` is the assessor with the largest proportion of times the test products were chosen.

Assessor `9` is among the assessors furthest to the left for both attributes. This means that assessor `9` tend to choose the control more often than the test products.

3. Summary and Discussion

We have in this paper suggested a way to analyze data from a binary paired comparison. The analysis that is suggested is to handle the replications of assessors by including them in the model, thus obtaining a Thurstonian

mixed model.
When considering Thurstonian mixed models an important gain is that the hypothesis test of a significant product effect handles the replications correctly. In addition, d-prime values of products as well as assessors are obtained from a Thurstonian mixed model. The assessor specific d-prime values enable a way to get information about the panel. In the situation with a non-significant assessor-by-product interaction, hypothesis tests and d-prime values are well-defined and interpretable. When the assessor-by-product interaction is significant, further research is needed to define and interpret hypothesis tests as well as the d-prime values for the main effects of products and assessors.
Throughout the paper, an analysis has been made for each attribute separately. Future work could be to investigate the possibility to account for correlations between the attributes.
An interesting continuation of the work presented in this paper is to consider other types of paired comparisons e.g. as in Gabrielsen (2000, 2001) providing an alternative analysis to existing ways of analyzing such data like Bradley-Terry models (Bi, 2015; Cattelan, 2012). Bradley-Terry models can be analyzed in `R` (R Core Team, 2017) by e.g. the `BradleyTerry2` package (Turner & Firth, 2012) and the `prefmod` package (Hatzinger & Maier, 2017; Hatzinger & Dittrich, 2012).

Acknowledgments

The research that lead to this paper is funded by the Technical University of Denmark and Unilever U.K. Central Resources Limited. Unilever also provided the data that were used as an example of the analyses in this paper. Furthermore, the first author would like to thank Rebecca Evans for many nice and rewarding discussions.

Appendix A. Implementation in `R`

The aim of this appendix is to illustrate how the methodology suggested in this paper can be implemented in `R`. To illustrate this we simulate data since the data used as an ongoing example throughout the paper are confidential.
Thurstonian models can be fitted in `R` using the `sensR` package (Christensen & Brockhoff, 2017) as illustrated in Brockhoff & Christensen (2010). When random effects are included in the model we use the `glmer` function from the `R`-package `lme4` (Bates et al., 2015).

A.1. Simulated data

The simulated data consist of the same variables as the data used in the paper. Thus, we are considering 8 products, 25 assessors and 2 sessions. The realizations of the response variable Y_{ijk} are simulated from the binomial distribution binomial$(1, 1/2)$.
Let `dat` be a data frame with a row for each observation Y_{ijk} with i, j and k representing the products, assessors and sessions respectively. Moreover, let the columns of `dat` be the response variable as well as the explanatory variables. More specifically, let `response` be the response variable and let `assessor`, `product` and `session` be the explanatory variables included here. `session` is not used in the model since we are considering models with effects of assessors and products.

A.2. R-code for the Thurstonian mixed model introduced in Section 2.4 using simulated data

When fitting generalized linear (mixed) models in `R` the link function must be specified. Throughout the paper the models have been written as:

$$p_{ij} = f_{psy}(\eta_{ij}) \tag{A.1}$$

Rewriting (A.1) makes it possible to identify how the link function is defined for these models:

$$\eta_{ij} = f_{psy}^{-1}(p_{ij})$$

Recall that the inverse of the psychometric function defined in (2) reads:

$$f_{psy}^{-1}(p_{ij}) = \Phi^{-1}(p_{ij})\sqrt{2} \tag{A.2}$$

This is the probit link function multiplied by the square root of 2. Thus, when fitting the models in R the probit link is used.

To make the functions from the `lme4` package available, simply write:

```
> library(lme4)
```

We obtain the Thurstonian mixed model in (6) with:

```
> (fm <-
+   glmer(response ~ product + (1|assessor) + (1|assessor:product),
+         data = dat,
+         family = binomial(probit),
+         contrasts = list("product"=contr.sum),
+         control=glmerControl(optimizer="bobyqa")))
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( probit )
Formula: response ~ product + (1 | assessor) + (1 | assessor:product)
   Data: dat
      AIC       BIC    logLik  deviance  df.resid
 562.4375  602.3522 -271.2188  542.4375       390
Random effects:
 Groups           Name        Std.Dev.
 assessor:product (Intercept) 0.1422
 assessor         (Intercept) 0.1992
Number of obs: 400, groups:  assessor:product, 200; assessor, 25
Fixed Effects:
(Intercept)     product1     product2     product3     product4     product5
   0.099290     0.268660    -0.202834    -0.099273     0.005319     0.005061
   product6     product7
  -0.099278     0.325808
```

where the `family` option is set to be `binomial(probit)` which means that the inbuilt link function `probit` is used for binomially distributed data.

To be able to do the hypothesis test of a significant assessor-by-product interaction, the model without the assessor-by-product interaction must be fitted:

```
> fm2 <-
+   glmer(response ~ product + (1|assessor),
+         data = dat,
+         family = binomial(probit),
+         contrasts = list("product"=contr.sum),
+         control=glmerControl(optimizer="bobyqa"))
```

We get the likelihood ratio test for the assessor-by-product interaction with:

```
> (LRT <- 2*(logLik(fm)[1] - logLik(fm2)[1]))
[1] 0.05112213
```

and the p-value is found as:

```
> (pVal <- 1 - pchisq(LRT, df = 1))
[1] 0.8211221
```

To obtain the estimates on the d-prime scale the estimates from the fitted model (`fm2`) must be multiplied by $\sqrt{2}$:

```
> (mu <- fixef(fm2)[1]*sqrt(2))
(Intercept)
      0.139
> alphas <- fixef(fm2)[-1]*sqrt(2)
> alpha.8 <- 0-sum(alphas)
> names(alpha.8) <- "product8"
> (alphas <- c(alphas, alpha.8))
 product1  product2  product3  product4
 0.376087 -0.283513 -0.138911  0.006969
 product5  product6  product7  product8
 0.006852 -0.138837  0.455908 -0.284554
> (bjs <- t(ranef(fm2)$"assessor"*sqrt(2)))
                    1       2      3
(Intercept) -0.04086 0.09038 0.2257
                    4      5      6
(Intercept) -0.2392 0.1588 0.1586
                    7      8       9
(Intercept) -0.04023 0.1583 0.09083
                   10      11      12
(Intercept) -0.2388 -0.1727 -0.1057
```

```
                13      14      15
(Intercept) -0.1721 -0.1056 0.09409
                16      17      18
(Intercept) 0.02763 0.3612 -0.03988
                19      20       21
(Intercept) -0.2391 0.02617 -0.04093
                22      23      24
(Intercept) 0.02478 -0.1725 0.02424
                25
(Intercept) 0.1567
```

where the value of α_8 is found using the restriction that the sum of $\alpha_1, \ldots, \alpha_8$ must equal zero.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bi, J. (2015). *Sensory Discrimination Tests and Measurements: Sensometrics in Sensory Evaluation*. Wiley Blackwell Publishing.

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

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical science*, *27*, 412–433.

Christensen, R. H. B., & Brockhoff, P. B. (2013). Analysis of sensory ratings data with cumulative link models. *Journal de la Société Francaise de Statistique*, *154*, 58–79.

Christensen, R. H. B., & Brockhoff, P. B. (2017). sensr—an r-package for sensory discrimination. R package version 1.5-0. http://www.cran.r-project.org/package=sensR/.

Christensen, R. H. B., Lee, H.-S., & Brockhoff, P. B. (2012). Estimation of the thurstonian model for the 2-ac protocol. *Food Quality and Preference*, *24*, 119–128.

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies*, *8*, 353–370.

Ennis, J. M. (2012). Guiding the switch from triangle testing to tetrad testing. *Journal of Sensory Studies*, *27*, 223–231.

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

Gabrielsen, G. (2000). Paired comparisons and designed experiments. *Food Quality and Preference*, *11*, 55–61.

Gabrielsen, G. (2001). A multi-level model for preferences. *Food Quality and Preference*, *12*, 337–344.

Hatzinger, R., & Dittrich, R. (2012). prefmod: An R Package for Modeling Preferences Based on Paired Comparisons, Rankings, or Ratings. *Journal of Statistical Software*, *48*, 1–31. URL: `http://www.jstatsoft.org/v48/i10/`.

Hatzinger, R., & Maier, M. J. (2017). *prefmod: Utilities to Fit Paired Comparison Models for Preferences*. URL: `https://CRAN.R-project.org/package=prefmod` R package version 0.8-34.

Næs, T., Brockhoff, P. B., & Tomic, O. (2010). *Statistics for Sensory and Consumer Science*. Wiley.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `https://www.R-project.org/`.

Turner, H., & Firth, D. (2012). Bradley-terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, *48*, 1–21. URL: `http://www.jstatsoft.org/v48/i09/`.

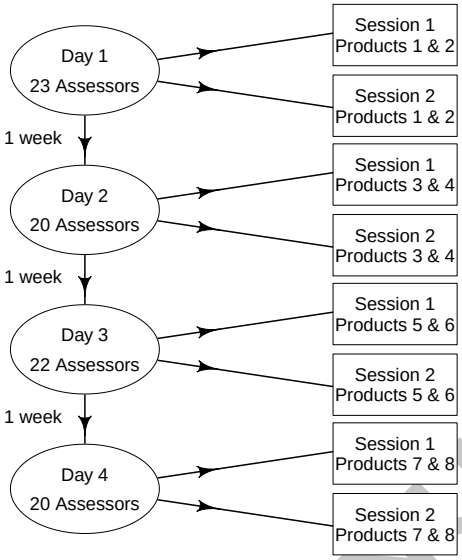


Figure 1: Organization of days, sessions, assessors as well as test products.

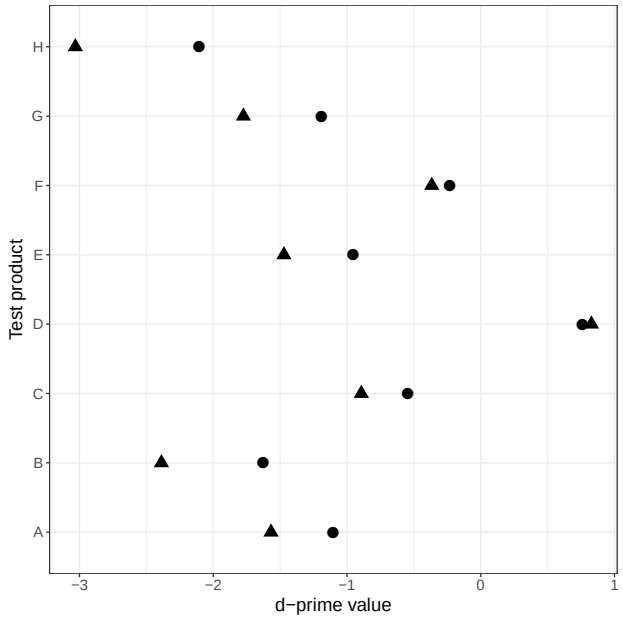


Figure 2: The d-prime values for test products for `Silky` after 5 minutes for model (3) (circles) and model (5) (triangles).

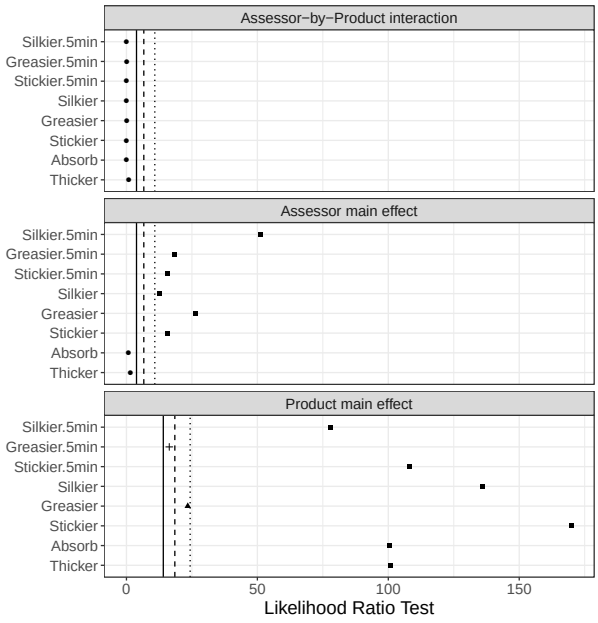


Figure 3: Likelihood Ratio Test statistics for the test of test products, the assessor main effect as well as the assessor-by-product interaction. The vertical lines are critical values for the corresponding Chi-squared distribution; the 0.05 critical value (full line), the 0.01 critical value (dashed line) and the 0.001 critical value (dotted line). The symbol shows the size of the corresponding p-value; a p-value that is less than 0.001 (square), a p-value between 0.001 and 0.01 (triangle), a p-value between 0.01 and 0.05 (plus) or a p-value larger than 0.05 (dot).

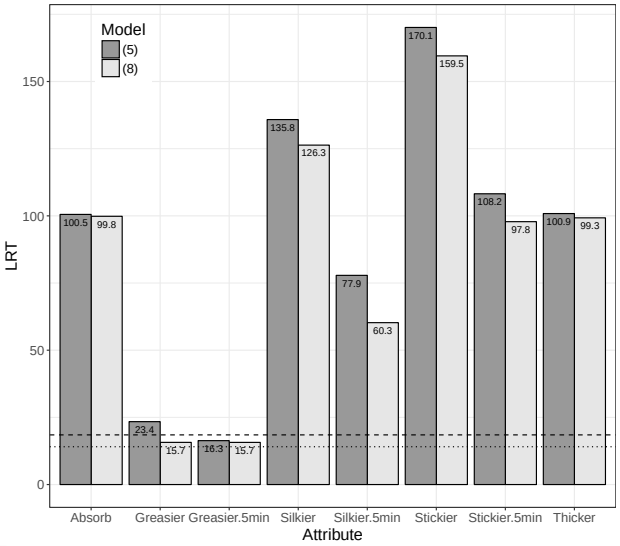


Figure 4: Comparing the likelihood ratio test statistics for hypothesis test of product main effect.The horizontal lines are the critical values for the Chi-squared distribution with 7 degrees of freedom; the 0.05 critical value (dotted line) and the 0.01 critical value (dashed line).

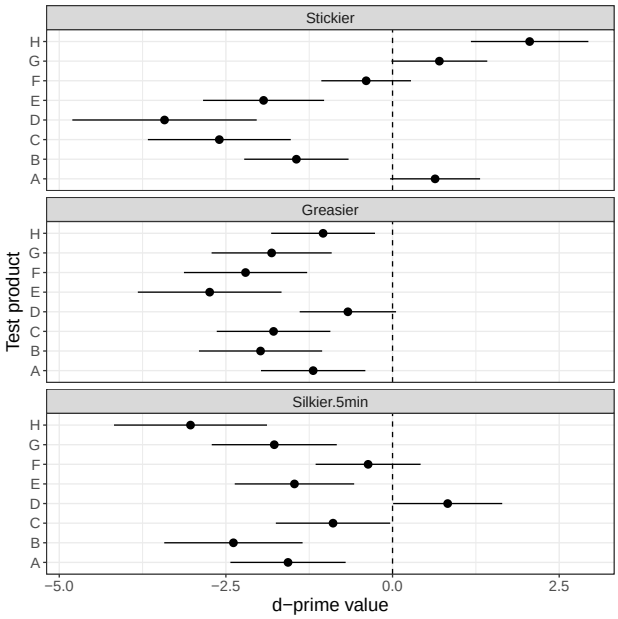


Figure 5: The d-prime estimates for the test products as well as 95% confidence intervals.

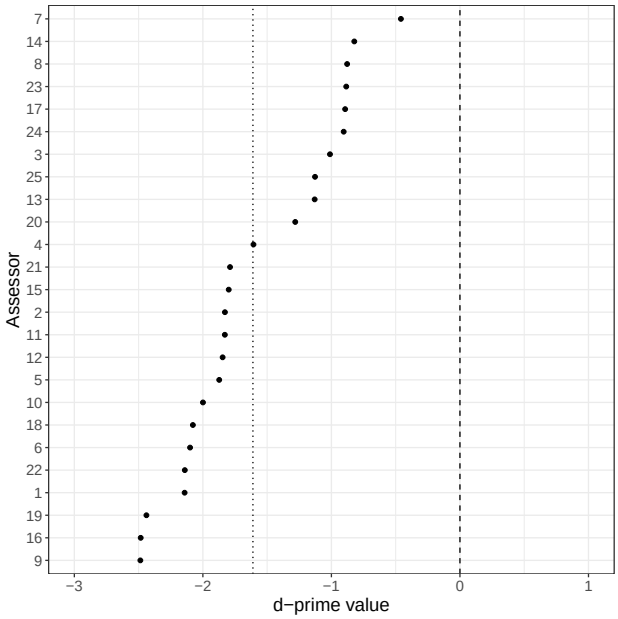


Figure 6: The sorted d-prime estimates of $\tilde{b}_j$ for the `Silky` attribute (0 minutes). The dotted line is the value of the consensus; the estimate of μ.

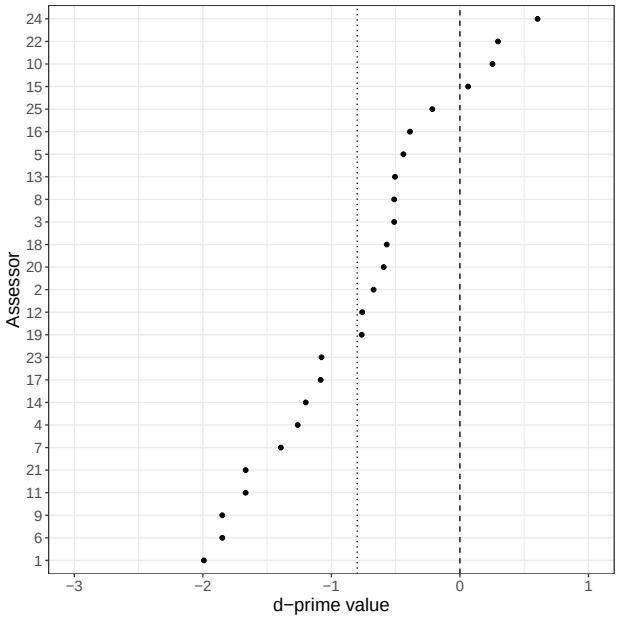


Figure 7: The sorted d-prime estimates of $\tilde{b}_j$ for the `Sticky` (0 minutes). The dotted line is the value of the consensus; the estimate of μ.

Table 1: Overview of the attributes.

Attribute	Evaluated after			
	0 min	5 min	Tactile	Visual
`Thickness`	✓		✓	
`Absorption`	✓		✓	✓
`Greasy`	✓	✓	✓	✓
`Sticky`	✓	✓	✓	
`Silky`	✓	✓	✓	

Table 2: The number of times a test product was chosen as having higher sensory intensity than the control for the eight attributes in percentages. The number of evaluations for the test products range from 40 to 46.

Test Product	Sticky		Greasy		Silky		Thickness	Absorption
	0 min	5 min	0 min	5 min	0 min	5 min	0 min	0 min
A	65.2	19.6	26.1	13.0	23.9	21.7	97.8	17.4
B	20.0	52.5	15.0	40.0	20.0	12.5	65.0	85.0
C	5.0	10.0	17.5	25.0	50.0	35.0	47.5	62.5
D	2.3	2.3	34.1	34.1	93.2	70.5	50.0	72.7
E	12.5	10.0	7.5	15.0	27.5	25.0	50.0	70.0
F	41.3	13.0	10.9	21.7	34.8	43.5	47.8	37.0
G	67.5	50.0	17.5	25.0	12.5	20.0	95.0	75.0
H	90.9	70.5	27.3	13.6	0.0	6.8	22.7	13.6

Table 3: d-prime values found by using the psychometric function on the proportions.

Test Product	Sticky		Greasy		Silky		Thickness	Absorption
	0 min	5 min	0 min	5 min	0 min	5 min	0 min	0 min
A	0.55	-1.21	-0.91	-1.59	-1.00	-1.10	2.86	-1.33
B	-1.19	0.09	-1.47	-0.36	-1.19	-1.63	0.54	1.47
C	-2.33	-1.81	-1.32	-0.95	0.00	-0.54	-0.09	0.45
D	-2.83	-2.83	-0.58	-0.58	2.11	0.76	0.00	0.86
E	-1.63	-1.81	-2.04	-1.47	-0.85	-0.95	0.00	0.74
F	-0.31	-1.59	-1.74	-1.10	-0.55	-0.23	-0.08	-0.47
G	0.64	0.00	-1.32	-0.95	-1.63	-1.19	2.33	0.95
H	1.89	0.76	-0.86	-1.55	-Inf	-2.11	-1.06	-1.55

Highlights:

Accounting for replications of assessors within sensory discrimination testing.

Replications of assessors are handled with a random assessor effect in the model.

Introducing how to obtain estimates of the assessors from the suggested model.

Enabling an analysis of multiple products accounting for replicates of the assessors.

Embedding the model into a Thurstonian framework leading to d-prime estimates.