Journal de la Société Française de Statistique
Vol. 154 No. 3 (2013)

Analysis of sensory ratings data with cumulative link models

Titre: Analyse des observations des évaluations sensorielles avec cumulative link models

Rune Haubo Bojesen Christensen[1] and Per Bruun Brockhoff[1]

Abstract: Examples of categorical rating scales include discrete preference, liking and hedonic rating scales. Data obtained on these scales are often analyzed with normal linear regression methods or with omnibus Pearson χ^2 tests. In this paper we propose to use cumulative link models that allow for regression methods similar to linear models while respecting the categorical nature of the observations. We describe how cumulative link models are related to the omnibus χ^2 tests and how they can lead to more powerful tests in the non-replicated setting. For replicated categorical ratings data we present a quasi-likelihood approach and a mixed effects approach both being extensions of cumulative link models. We contrast population-average and subject-specific interpretations based on these models and discuss how different approaches lead to different tests. In replicated settings, naive tests that ignore replications are often expected to be too liberal because of over-dispersion. We describe how this depends on whether the experimental design is fully randomized or blocked. For the latter situation we describe how naive tests can be stronger than over-dispersion adjusting approaches, and that mixed effects models can provide even stronger tests than naive tests. Examples will be given throughout the paper and the methodology is implemented in the authors' free R-package ordinal.

Résumé : Les données issues d'une étude hédonique ou de préférence sont généralement représentées avec une échelle à catégories ordonnées. Elles sont souvent analysées par des méthodes de régression linéaire ou des tests omnibus de Khi-deux de Pearson. Nous proposons dans cet article le recours à des modèles de régression à fonction de lien cumulée qui respectent la nature ordinale des observations. Nous décrivons comment ces modèles sont liés aux tests omnibus de Khi-deux, et comment ils peuvent conduire à des tests plus puissants en l'absence de répétitions. Pour les notations sur une échelle ordinale, nous présentons une approche de type maximum de quasi-vraisemblance et une approche de type "modèles mixtes" qui sont en fait des extensions du modèle à fonction de lien cumulée. Avec ces modèles nous comparons les interprétations de l'effet moyen et de l'effet spécifique du sujet, et nous discutons comment les différentes approches conduisent à différents tests. En présence de répétitions, les tests «naïfs» qui ignorent celles-ci sont souvent trop permissifs à cause de la sur-dispersion. Nous discutons aussi de la dépendance du plan expérimental, randomisé ou en blocs. Pour ces plans en blocs nous abordons la question de savoir comment les tests naïfs peuvent être plus puissants que les approches qui prennent en compte la sur-dispersion et comment les modèles mixtes peuvent fournir des tests encore plus puissants que des tests naïfs. Des exemples sont présentés tout au long de l'article. Les procédures d'analyse sont implémentées par les auteurs dans l'environnement R.

Keywords: Cumulative link models, ordinal regression models, mixed effects models, R software
Mots-clés : modèle à fonction de lien cumulée, modèle de régression ordinale, modèle mixte, logiciel R
AMS 2000 subject classifications: 62H17, 62J12

[1] DTU Informatics, Statistical section, Technical University of Denmark, Richard Petersens Plads, Building 305, DK-2800 Kongens Lyngby, Denmark.
E-mail: rhbc@imm.dtu.dk

 ISSN: 2102-6238

1. Introduction

By categorical ratings data we mean data observed on an ordered categorical scale with at least two categories. This includes the common 5, 7, and 9 points preference, liking and hedonic rating scales, but excludes finite continuous scales as are used in sensory profiling. The categorical rating scales are common in sensory science as well as many other sciences where humans are used as measurement instruments (Greene and Hensher, 2010).

There are often clear grouping structures in such data because each subject provides several observations — a concept that is known in the sensometric literature as *replications*. Since two observations from the same individual are likely to be more similar on average than observations from different individuals, the observations are not independent and conventional statistical tests no longer apply directly. The main objective of this paper is to propose statistical tests and models for categorical ratings data that handle grouping structures in the data appropriately. The approach we consider here is based on cumulative link models (CLMs); a well-known class of statistical models (McCullagh, 1980; Agresti, 1999, 2002; Greene and Hensher, 2010).

A simple approach often described in introductory text books is to use normal linear models (regression and ANOVA) directly on the ratings under equal distance numbering of the categories. This approach can be a useful approximation if there are sufficiently many categories and not too many observations in the end categories, but it treats inherently categorical data as continuous. It is hard to quantify how this affects accuracy and consistency of parameter estimates as well as testing accuracy and power. In particular for rating scales with a small number of categories, linear models are inappropriate. A more appealing approach is to treat the observations rightfully as categorical as we do in this paper.

The conventional omnibus χ^2-statistics treat data as categorical, but they do not utilize the ordering of the categories. In section 2 it will be described how cumulative link models utilize this ordering and that they often lead to stronger tests than the omnibus tests.

Tests for replicated categorical data were considered by Ennis and Bi (1999), who proposed the Dirichlet-Multinomial (DM) model. Conceptually this model is equivalent to the beta-binomial model (Ennis and Bi, 1998; Brockhoff, 2003) for multinomial rather than binomial observations. The idea is to adjust conventional statistical tests for over-dispersion. Although the DM model is applicable to ordinal data, it does not take advantage of the ordered nature of the observations.

The first approach to handling replications in categorical ratings data that we discuss is akin to the DM model in that it adjusts standard errors for over-dispersion. The amount of over-dispersion is estimated in a quasi-likelihood framework for cumulative link models. In contrast to the DM model, this approach respects the ordinal nature of the observations.

The second approach to handling replications that we propose is based on cumulative link mixed models (CLMMs) which include random effects for the grouping variable (Agresti and Natarajan, 2001). Conceptually this is an extension of linear mixed models to ordinal observations, but computationally this model class turns out to be much more complicated. Model specification and interpretation also turns out to be more complex partly due to the discrete nature of the observations and partly due to the fact that the model is nonlinear in its parameters. Due to the nonlinearity of the link function, the two approaches that we propose lead to different interpretations. The mixed models have so-called *subject-specific* interpretations while the over-dispersion adjusted models have *population-average* interpretations. The quasi-likelihood approach is a simple alternative to

 ISSN: 2102-6238

the more satisfying, but also more complicated, framework of cumulative link mixed models.

In section 2 we outline cumulative link models, we describe their relation to standard omnibus χ^2 tests and the advantages of cumulative link models over these tests. We also describe a latent variable interpretation of cumulative link models that connects these with Thurstonian models. In the context of sensory discrimination tests, Thurstonian models are stochastic descriptions of sensory perception. They provide a general measure of sensory difference as the mean difference between latent normal distributions (Thurstone, 1927a; Lawless and Heymann, 1998). In section 3 we describe a quasi-likelihood approach to handle replicated ratings data and describe similarities and differences to the DM model. In section 4 we describe cumulative link mixed models for replicated ratings data and contrast this approach to the quasi-likelihood approach and the DM model. Most emphasis is given to the approach of cumulative link mixed models because we find that this gives the most appealing and flexible framework for modeling replicated ratings data. We end with discussions in section 5. Examples are given throughout the paper illustrating the different approaches on data from the literature. These datasets for our examplescan be read of from tables in the original publications. A software implementation of the methodology described in this paper is available in the R-package ordinal (Christensen, 2012) developed by the authors freely available for the statistical software R (R Development Core Team, 2010).

2. Cumulative link models for non-replicated ratings data

In this section we outline standard cumulative link models that do not account for replications. We describe how association, e.g. product differences, can be tested in CLMs and we establish the connection to the conventional χ^2-statistics. We also describe an appealing latent variable interpretation of CLMs.

2.1. Outline of cumulative link models

A cumulative link model for an ordinal variable, Y_i that can fall in J categories is a linear model for a transformation of cumulative probabilities, γ_{ij} through a link function:

$$P(Y_i \leq j) = \gamma_{ij} = F(\theta_j - \boldsymbol{x}_i^T \boldsymbol{\beta}) \quad i = 1,\ldots,n \quad j = 1,\ldots,J \tag{1}$$

where the intercept parameters

$$-\infty \equiv \theta_0 \leq \theta_1 \leq \ldots \leq \theta_{J-1} \leq \theta_J \equiv \infty \tag{2}$$

are ordered, F is the so-called inverse link function and $\boldsymbol{x}_i^T$ is a p-vector of regression variables for the parameters, $\boldsymbol{\beta}$. The linear model, $\boldsymbol{x}_i^T \boldsymbol{\beta}$ is assumed to apply in the same way across all response categories as it does not depend on j. A typical choice of link function is the probit link, $F^{-1} = \Phi^{-1}$, where Φ is the standard normal cumulative distribution function. We will adopt this choice throughout and motivate it in section 2.5. While the linear model, $\boldsymbol{x}_i^T \boldsymbol{\beta}$ is known as the location structure, the cumulative link model may also be extended with a scale structure, $\exp(\boldsymbol{z}_i^T \boldsymbol{\zeta})$ so that the resulting location-scale cumulative link model cf. Cox (1995); Agresti (2002); Christensen et al. (2011) reads

$$\gamma_{ij} = F\left(\frac{\theta_j - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\exp(\boldsymbol{z}_i^T \boldsymbol{\zeta})}\right) \quad i = 1,\ldots,n \quad j = 1,\ldots,J \tag{3}$$

 ISSN: 2102-6238

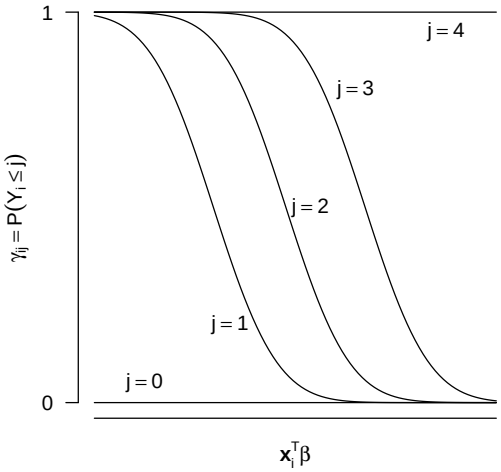

FIGURE 1. *Illustration of a cumulative link model with four response categories.*

The cumulative link model (1) is illustrated in Fig. 1 where $F = \Phi$ and $J = 4$ is adopted. The horizontal displacement of the three curves is determined by the values of θ_j for $j = 1,\ldots,J-1$. The cumulative probabilities of an observation falling in each of the response categories can be read of the vertical axis for a value of the linear model, $\boldsymbol{x}_i^T\boldsymbol{\beta}$. The lines for $j = 0$ and $j = 4$ are horizontal straight lines at 0 and 1 by definition.

The ordinal response variable, Y_i can be represented by the vector $\boldsymbol{Y}_i^* = (Y_{i1}^*,\ldots,Y_{ij}^*,\ldots,Y_{iJ}^*)$ where $Y_{ij}^* = 1$ if Y_i falls in the jth category, i.e. if $Y_i = j$ is observed and zero otherwise. $\boldsymbol{Y}_i^*$ is said to follow the multinomial distribution $\boldsymbol{Y}_i^* \sim \text{multinom}(1,\boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i$ is the probability parameter vector for the ith observation with elements $\pi_{ij} = P(Y_i = j) = P(Y_{ij}^* = 1)$. The parameters satisfy $\sum_{j=1}^J \pi_{ij} = 1$ and are linked to the cumulative probabilities by $\gamma_{ij} = \sum_{h=1}^j \pi_{ih}$, or equivalently $\pi_{ij} = \gamma_{ij} - \gamma_{i,j-1}$.

The probability mass function for this multinomial distribution is the multivariate extension of the Bernoulli probability mass function $P(\boldsymbol{Y}^* = \boldsymbol{y}^*) = \prod_{i=1}^n \prod_{j=1}^J \pi_{ij}^{y_{ij}^*}$, so the log-likelihood function can be expressed as

$$\ell(\boldsymbol{\alpha};\boldsymbol{y}) = \sum_{i=1}^n w_i \sum_{j=1}^J y_{ij}^* \log \pi_{ij}$$

where w_i is a potential weight for the ith observation and $\boldsymbol{\alpha}$ is a vector of all parameters.

2.2. *Testing in cumulative link models*

In this section approaches to tests of association in cumulative link models are outlined. We will consider the situation in which $k = 1,\ldots,K$, $K \geq 2$ products are rated on an ordinal scale with $j = 1,\ldots,J$, $J \geq 2$ categories with respect to preference, liking or some other aspect of interest. The objective is to assess if and how ratings differ among products. We will assume that k index rows and j index columns in the resulting two-way multinomial table.

 ISSN: 2102-6238

Tests of association in these two-way multinomial tables can be done via likelihood ratio tests. The likelihood ratio statistic is $LR = -2\{\ell_0(\hat{\boldsymbol{\alpha}};\boldsymbol{y}) - \ell_1(\hat{\boldsymbol{\alpha}};\boldsymbol{y})\}$ for the comparison of two nested models m_0 and m_1 and $\hat{\boldsymbol{\alpha}}$ is the ML estimates under the models. The likelihood ratio statistic asymptotically follows a χ^2-distribution with degrees of freedom equal to the difference in the number of parameters for the models being compared. For binomial and multinomial observations, this statistic can also be expressed as

$$G^2 = 2\sum_{k,j} e_{1kj} \log \frac{e_{1kj}}{e_{0kj}} \tag{4}$$

where e_{1kj} and e_{0kj} are the expected counts under models m_1 and m_0 (Agresti, 2002; McCullagh and Nelder, 1989). For ordinal data each row in the table is a multinomial vector which has its sum fixed by design. The expected counts are therefore given by $e_{kj} = \pi_{kj} r_k$, where π_{kj} is the fitted probability in cell (kj) and $r_k = \sum_j o_{kj}$ is the sum of the observed counts in row k. A closely related and often very similar statistic (Agresti, 2002; McCullagh and Nelder, 1989) is Pearson's statistic:

$$X^2 = \sum_{k,j} \frac{(e_{0kj} - e_{1kj})^2}{e_{0kj}} \tag{5}$$

These two statistics measure the discrepancy between the models m_1 and m_0 and are related through the power-divergence family (Cressie and Read, 1989).

Another statistic which is generally inferior to these two statistics is the Wald statistic (Pawitan, 2000, 2001). To test the existence of an effect described by a parameter vector, $\boldsymbol{\alpha}$ of length p, the multivariate Wald statistic (Wasserman, 2004) reads

$$W = \hat{\boldsymbol{\alpha}}^T \mathrm{Cov}(\hat{\boldsymbol{\alpha}})^{-1} \hat{\boldsymbol{\alpha}} \tag{6}$$

which follows asymptotically a χ^2-distribution with p degrees of freedom under the null hypothesis and $\mathrm{Cov}(\hat{\boldsymbol{\alpha}})$ is the variance-covariance matrix of the parameters at their maximum likelihood estimates. For scalar $\boldsymbol{\alpha}$ this can be simplified to $\sqrt{W} = \hat{\alpha}/\mathrm{se}(\hat{\alpha})$ which follows asymptotically a standard normal distribution.

2.3. *Connection to conventional χ^2 statistics*

In this section the connection between testing in cumulative link models and conventional omnibus G^2 and X^2 tests is explored. The omnibus χ^2-tests can be written as in eq. (4) and eq. (5), where e_1 are the observed cell counts and e_0 are the expected cell counts given by the familiar formula $e_{0kj} = r_k \cdot c_j / N$, where $c_j = \sum_k e_{1kj}$ are the column totals and $N = \sum_{kj} e_{1kj}$ is the overall sum. The statistics asymptotically follow a χ^2-distribution on $(J-1)\cdot(K-1)$ degrees of freedom. Formally, the omnibus tests assume the following null and alternative hypotheses for our setting:

$$\begin{aligned} H_0 &: \quad \boldsymbol{Y}_k^* \sim \mathrm{multinom}(m_k, \boldsymbol{\pi}) \\ H_1 &: \quad \boldsymbol{Y}_k^* \sim \mathrm{multinom}(m_k, \boldsymbol{\pi}_k) \end{aligned} \tag{7}$$

where H_0 specifies that the multinomial probability does not depend on k with $J-1$ parameters, and H_1 specifies that the multinomial probability depends on k with $(J-1)\cdot K$ parameters. The

 ISSN: 2102-6238

difference in the number of parameters is $(J-1)\cdot(K-1)$, further, the expected counts under model m_1 corresponding to H_1 are exactly the observed counts, so testing the hypotheses in (7) is equivalent to application of the omnibus G^2 and X^2 tests.

The models implied by the hypotheses in (7) can be written as cumulative link models; $m_0: \ \gamma_j = \Phi(\theta_j)$ and $m_1: \ \gamma_{jk} = \Phi(\theta_{jk})$, where the cumulative probabilities, $\boldsymbol{\gamma}$ are linked to $\boldsymbol{\pi}$ as described in section 2.1. The model, m_0 implied by H_0 is known as the *null* model because it describes no other structure than that imposed by design, and model m_1 implied by H_1 is known as the *full* model because it completely describes the observed data with no residual degrees of freedom.

One of the main benefits of cumulative link models is that models intermediate between the null and full models can easily be specified and this often leads to stronger tests of product differences or other associations.

A cumulative link model that specifies a location difference, i.e. an additive shift on the probit scale reads

$$\gamma_{jk} = \Phi(\theta_j - c_k) \quad j = 1,\ldots,J \quad k = 1,\ldots,K \geq 2 \tag{8}$$

where c_k describes the effect of the kth product. This model uses $(J-1)+(K-1)$ degrees of freedom, which, for $J > 2$, is less than the full model given by H_1 and therefore a model intermediate to the null and full models.

A model that specifies location as well as scale differences, i.e. additive and multiplicative effects on the probit scale reads

$$\gamma_{jk} = \Phi\{(\theta_j - c_k)/g_k\} \quad j = 1,\ldots,J \quad k = 1,\ldots,K \geq 2 \tag{9}$$

where g_k is the multiplicative effect of the kth product. This model uses $(J-1)+2(K-1)$ degrees of freedom which is less than the full model if $J > 3$ and equal to the full model if $J = 3$. For $J \geq 4$ a comparison of model (9) to the full model can be considered a test of differences of higher order than location and scale differences. In general the comparison of a particular working model to the full model is a goodness-of-fit (GOF) test of that model. Recall that an insignificant GOF test does not imply that the model fits well, only that the test had not enough power to provide a significant result. On the other hand a model based on plenty of data can yield a significant GOF test while still being useful and possibly an appealing model for the data generating mechanism — consequently GOF tests may be used rather informally.

Usually differences of higher order than location and scale are hard to identify and often even scale differences are negligible. The discrepancies between location and null models will therefore often be comparable in size to the omnibus G^2 and X^2 statistics but on fewer degrees of freedom and therefore provides a more powerful test.

An approach related to cumulative link models is that of decomposition of χ^2 statistics. The basic idea is that the omnibus statistics can be decomposed into orthogonal components each having a χ^2 distribution such that all components with appropriate degrees of freedom add up to the omnibus test. One degree of freedom tests for location and scale differences can be constructed in this way. (Agresti, 2002, sec. 3.3.3) gives a brief description of the basic idea and Nair (1986) is a thorough description and discussion of a particular decomposition. Similar ideas are described by Rayner and Best (2001) and Rayner et al. (2005) and briefly considered in (Bi, 2006, sec. 5.3.2).

TABLE 1. *Ratings for a replicated paired degree-of-difference test adopted from Bi (2002). Data are aggregated over assessors.*

	rating[a]		
pair	1	2	3
concordant	45	40	15
discordant	36	34	30

[a]: 1 means identical and 3 means different

In contrast to the cumulative link models, the nonparametric approach does not easily generalize to the regression framework and to replicated data. While tests merely describe the degree of evidence of association, a model based approach also makes it possible to investigate the nature of association; the direction of differences and the strength of association, see (Agresti, 2002, sec. 3.3.6 and 3.4) for further discussion.

2.4. Example 1

In this example we compare various χ^2 tests. Bi (2002) describes a replicated paired degree-of-difference test where 25 subjects each assess four concordant and four discordant product pairs. The subjects were asked to rate the degree of difference between the sample product pairs on a three point rating scale, where 1 means *identical* and 3 means *different*. In this example we will ignore the grouping structure in the data and analyze the data as if they were independent. The data are summarized in Table 1.

A test of differences in ratings between concordant and discordant sample pairs is a test of product differences. Using eq. (5), we find that the omnibus Pearson χ^2-test statistic is $X^2 = 6.49$. On 2 degrees of freedom, we may find using tables or statistical software that the p-value is $p = 0.039$. Similarly, by application of eq. (4) the omnibus likelihood ratio χ^2-test statistic is $G^2 = 6.59$, which on df = 2 gives $p = 0.037$. Since the full model is equivalent to the location-scale model (9), the same test could be obtained as a likelihood ratio test of the comparison of model (9) and the null model.

The likelihood ratio test of a location difference is obtained by comparing the null model with model (8). This leads to $G^2 = 4.70$, df = 1, $p = 0.030$ and therefore a slightly stronger test than the omnibus tests. The likelihood ratio test of scale and higher order differences while controlling for location differences is $G^2 = 1.88$, df = 1, $p = 0.170$. This test can be obtained as the likelihood ratio test of models (8) and (9) or, since the χ^2 statistics are additive, as the difference between the omnibus G^2 test and the likelihood ratio test of a location difference: $6.59 - 4.70 = 1.88$ save for rounding errors. Observe also that the likelihood ratio test of scale and higher order differences can be considered a GOF test of the location model (8).

The main discrepancy in these data is due to location differences, while there is no evidence of differences in scale and higher order moments. The test of location differences is a stronger test than the omnibus tests because the main discrepancy in the table can be summarized as a location difference on only one degree of freedom. This is a fairly typical result that is often even more pronounced in situations with more response categories.

 ISSN: 2102-6238

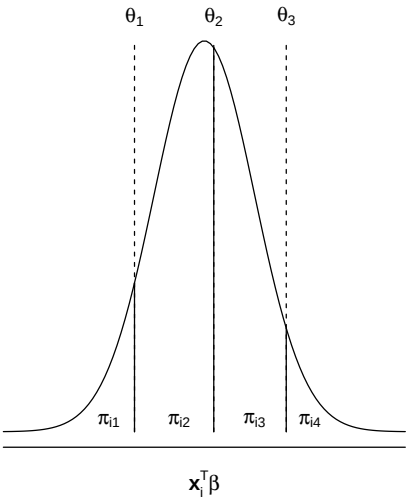

FIGURE 2. *Illustration of a cumulative link model in terms of the latent distribution.*

2.5. Latent variable interpretation

The cumulative link model can be interpreted as a model for a continuous latent variable. Suppose for instance that preference for a particular product type, S can be described by a normal linear model: $S \sim N(\mu, \sigma^2)$, where μ describes structural differences in preference, for instance the average difference in preference between consumers from different regions and σ is the residual standard deviation. The variation in preference could be due to differences in product samples, differences in perception of the samples or variations in preference. Preference, S is not observed directly — only a categorized version, Y is observed. This latent variable interpretation is conceptually similar to the Thurstonian model of paired preferences (Thurstone, 1927a,b,c). Suppose that Y is observed in the jth category if S falls between the thresholds θ_{j-1} and θ_j obeying (2), then the cumulative probabilities can be expressed as a function of the model parameters: $\gamma_{ij} = \Phi(\theta_j - \mu_i)$. This is the cumulative link model with a probit link, where μ_i can be described by a general linear predictor; $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ as in eq. (1). In this model μ_i refers to a location difference relative to the origin, μ^0 and scale, σ of the latent distribution; $\mu_i = (\mu_i^* - \mu^0)/\sigma$; similar arguments appear in Catalano and Ryan (1992) and Christensen et al. (2011).

The latent variable interpretation of a cumulative link model is illustrated in Fig. 2 where a probit link and $J = 4$ is adopted. The three thresholds, $\theta_1, \ldots, \theta_3$ divide the area under the curve into four parts each of which represent the probability of a response falling in the four response categories. The thresholds are fixed on the scale, but the location of the latent distribution, and therefore also the four areas under the curve, change with $\boldsymbol{x}_i^T \boldsymbol{\beta}$. Assuming other latent distributions lead to other link functions, for example, assuming that S has a logistic distribution leads to a logit link. The location-scale model (cf. eq. (3) and (9)) arise if the spread of the latent distribution is also allowed to depend on i.

Likelihood ratio tests of effects are often fairly unaffected by the choice of link function and often very large amounts of data are necessary to distinguish between the links in terms of goodness-of-fit (Genter and Farewell, 1985). Different link functions, however, lead to different

 ISSN: 2102-6238

parameter estimates and interpretations can differ.

2.6. Example 2

The test of location differences in example 1, section 2.4, is a test of $H_0: \ c_2 - c_1 = 0$ versus $H_1: \ c_2 - c_1 \neq 0$ in model (8) with $K = 2$ and $J = 3$. This implies the latent distributions; $S_k \sim N(\mu^0 + c_k, \sigma)$, where the absolute location, μ^0 and scale σ are unknown and not estimable from data, but the maximum likelihood estimate of the location difference $c_2 - c_1$ is 0.3462 with standard error 0.160. The maximum likelihood estimates of the threshold parameters are $\hat{\boldsymbol{\theta}} = (-0.073, 0.939)$.

3. Adjusting for over-dispersion in replicated ratings data

It was recognized by Ennis and Bi (1999) that standard statistical tests are not appropriate when grouping structures violate the assumption of independent observations. They proposed to adjust the test statistics by an amount related to the degree of over-dispersion in the data relative to what would be expected for independent observations. The degree of over-dispersion is estimated in a Dirichlet-multinomial (DM) model where the multinomial probabilities are allowed to vary. A disadvantage of this approach for rating data is that it treats ordinal data as unordered. We present in this section an alternative approach of adjusting tests for over-dispersion within models that take advantage of the ordered nature of the ratings.

3.1. Quasi-likelihood approach

A well-known way of modeling over-dispersed discrete data is the quasi-likelihood approach (Wedderburn, 1974; McCullagh and Nelder, 1989). The basic idea is to model the population mean of the observations with a linear predictor through a link function. The amount of over-dispersion is estimated by comparing the observed variation with the variation that would be expected if the observations were independent. This approach conceptually amounts to estimating parameters with a standard cumulative link model, but adjusting the variance-covariance matrix of the parameter estimates by multiplication with an over-dispersion parameter ϕ.

Over-dispersed cumulative link models are specified in terms of the first two moments of the distribution of the observations and so does not assume a full likelihood specification including higher order moments. This means that likelihood ratio tests are unavailable, but Wald tests of individual parameters and multivariate Wald tests of model terms can be constructed. Approximate F-tests for model terms can also be constructed that are similar to likelihood ratio tests for likelihood based models, but these F-tests tend to be rather conservative, so we do not consider them further. See Collett (2002) for construction of these F-tests in binomial models; see also McCullagh and Nelder (1989) and Venables and Ripley (2002) for relevant discussion. In this approach, observations are assumed to follow a quasi-multinomial distribution with expectation $\mathsf{E}[\boldsymbol{Y}] = m\boldsymbol{\pi}$ and covariance $\mathsf{Cov}[\boldsymbol{Y}] = \phi m(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)$, where the over-dispersion parameter, ϕ distinguishes the distribution from a genuine multinomial distribution which has $\phi \equiv 1$. A cumulative link model is assumed to describe the mean structure in the observations and the

 ISSN: 2102-6238

TABLE 2. *Replicated degree of liking data from 104 subjects reported by Ennis and Bi (1999) aggregated over subjects.*

	rating[a]				
city	1	2	3	4	5
New York	9	26	15	120	154
San Francisco	5	21	28	129	117

[a]: 1 means "dislike very much" and 5 means "like very much".

variance-covariance matrix of the parameters is given by $\mathrm{Cov}[\boldsymbol{\alpha}] = \phi \boldsymbol{G}_{\boldsymbol{\alpha}}$, where $\boldsymbol{G}_{\boldsymbol{\alpha}}$ is the variance-covariance matrix of $\boldsymbol{\alpha}$ in a standard full likelihood approach. The quasi-likelihood approach therefore amounts to inflating the standard errors of the parameter estimates with, $\sqrt{\phi}$.

There are two standard approaches to estimate ϕ related to the G^2 and X^2 statistics (Pawitan, 2001; McCullagh and Nelder, 1989):

$$\hat{\phi}_G = \frac{2}{n-p} \sum_{k,j} e_{1kj} \log \frac{e_{1kj}}{e_{0kj}}$$

$$\hat{\phi}_P = \frac{1}{n-p} \sum_{k,j} \frac{(e_{0kj} - e_{1kj})^2}{e_{0kj}}$$

where $n = (J-1) \cdot K$ is the total number of degrees of freedom, p is the number of parameters in the model, e_{1kj} are the observed cell counts and e_{0kj} are the expected cell counts under the model. This corresponds to a generalized estimation equation (GEE) approach assuming a so-called independence working correlation model (Fahrmeir and Tutz, 2001, sec. 3.5). The estimators are only valid when the multinomial table is not sparse; as a general rule the expected frequencies should be at least five. There are generally only minor differences between the two ϕ-estimators. A considerable difference is an indication that the model is inappropriate and tests in the model should not be trusted. When this occurs the expected frequencies are small or important structures have been ignored in the data.

Under the quasi-likelihood model a modified Wald statistic, $W^* = W/\hat{\phi}$ is used instead of the standard Wald statistic (6) to test association.

3.2. Example 3

In Ennis and Bi (1999) the degree of liking among consumers in a replicated rating experiment conducted in New York and San Francisco was considered. A five-point liking scale was adopted where 1 means "dislike very much" and 5 means "like very much". 54 subjects from New York and 50 subjects from San Francisco were included in the study and each of the subjects evaluated six samples of the product. The main objective is to consider whether there is a difference in liking between cities. Data are summarized in Table 2 aggregated over subjects.

The omnibus Pearson and likelihood ratio tests applied directly to Table 2 yield $X^2 = 10.07$, df = 4, $p = 0.039$ and $G^2 = 10.15$, df = 4, $p = 0.038$ indicating a difference between the cities with respect to liking.

The joint test of location and scale differences while ignoring the grouping structure (replications) in the location-scale model (9) with $K = 2$ and $J = 5$ is $LR = 6.71$, df = 2, $p = 0.035$. The multivariate Wald test for the same hypothesis yields $W = 7.31$, df = 2, $p = 0.023$. The likelihood

 ISSN: 2102-6238

ratio test for higher order differences is $LR = 3.44$, df = 2, $p = 0.179$, so there is no evidence of more than location and scale differences.

In Ennis and Bi (1999) a Dirichlet-Multinomial (DM) model was fitted to the data from each of the two cities and obtain estimates of over-dispersion correction parameters, which employed in a (bivariate) Wald test yields p-value = 0.38 of the difference in liking between the two cities. They conclude that when adjusting for over-dispersion, there is no evidence of a difference in liking between the two cities.

The estimate of ϕ_G based on model (9) is the likelihood ratio statistic for the test of higher order differences scaled by the residual degrees of freedom, $n-p = 4\cdot 2-6 = 2$, i.e. $\hat{\phi}_G = 3.44/2 = 1.72$. The Wald statistic for the joint test of location and scale differences is $W^* = 7.31/1.72 = 4.25$, which on 2 degrees of freedom gives $p = 0.119$. This test is adjusted for the over-dispersion caused by the replications and is therefore more appropriate than the naive test; consequently the naive test assuming independent observations is too liberal. The test of regional differences based on quasi-likelihood leads to an answer ($p = 0.119$) in between the too liberal naive test ($p = 0.038$) and the DM model proposed by Ennis and Bi (1999) ($p = 0.38$).

4. Cumulative link mixed models for replicated ratings data

In this section we consider an extension of cumulative link models where random effects are included in the location part of the predictor. As such it can also be viewed as an extension of linear mixed models to ordered categorical observations. This framework is more flexible than the quasi-likelihood approach and allows for a more insightful modeling of grouping structures. Cumulative link mixed models is a member of a class of models sometimes referred to as multivariate generalized nonlinear mixed models (Fahrmeir and Tutz, 2001). The latent variable interpretation carries over to the mixed versions of cumulative link models and if the probit link is assumed, the model amounts to a standard linear mixed model for the latent variable. A cumulative link mixed model with a single random effect term can be expressed as

$$\gamma_{ijl} = F(\theta_j - \boldsymbol{x}_{il}^T\boldsymbol{\beta} - b_i) \quad i = 1,\ldots,n \quad l = 1,\ldots,l_i \quad j = 1,\ldots,J$$

where it is assumed that the conditional distribution of the observations given the realizations of the random effects is multinomial and the random effects are normally distributed

$$(Y_{il}|B_i = b_i) \sim \text{Multinom}(1, \boldsymbol{\pi}_{il}) \quad B_i \sim N(0, \sigma_b^2)$$

The $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ parameters describe the structure in the conditional distribution of the observations, and σ_b describes the heterogeneity in the population. This model is akin to a normal linear mixed model where the response is treated as ordinal rather than normally distributed. If the inverse link function, F is taken to be the standard normal cumulative distribution function, this model corresponds to assuming the following linear mixed model for the latent scale:

$$S_{il} = \boldsymbol{x}_{il}^T\boldsymbol{\beta} + b_i + e_{il} \quad E_{il} \sim N(0, \sigma^2) \quad B_i \sim N(0, \sigma_b^2) \tag{10}$$

This is possibly the simplest model for the latent scale that accounts for the grouping structure in the data.

 ISSN: 2102-6238

The population spread, σ_b is estimated relative to the spread of the latent scale, so it can be interpreted as a ratio of between-to-within subject variation. Observe also that the size of σ_b changes with the link function. This is not only because this means another mapping from the linear predictor to the probability scale, but primarily because the variance of the residuals (cf. eq. (10)) change with the distributional assumptions entailed by the link function. For instance a logit link corresponds to assuming a logistic distribution for the latent scale, and since the standard logistic distribution has variance $\pi^2/3$, the estimated σ_b will be scaled by approximately $\pi/\sqrt{3}$ compared to the estimate obtained with a probit link.

The log-likelihood function for the models we consider may be written as

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_b; y) = \sum_{i=1}^{n} \log \int_{\mathbb{R}} p(\boldsymbol{y}_i|b_i) p(b_i)\, \mathrm{d}b_i \tag{11}$$

where $p(\boldsymbol{y}_i|b_i)$ is the conditional multinomial probability mass function of the observations given the random effects, and $p(b_i)$ is the (marginal) normal density of the random effects. The log-likelihood is a sum of n independent contributions since observations from different individuals are assumed independent.

Estimation of CLMMs is complicated by the fact that the integral in eq. (11) does not have a closed form solution. Several different approximations have been proposed and two of the most popular are the Laplace approximation and adaptive Gauss-Hermite quadrature (AGQ) (Liu and Pierce, 1994; Pinheiro and Bates, 1995; Skrondal and Rabe-Hesketh, 2004; Joe, 2008). The Laplace approximation is a fast and reasonably accurate approximation while AGQ is computationally more intensive, but it has the advantage that the accuracy can be increased by adding more quadrature nodes. Often the Laplace approximation is sufficiently accurate while essentially exact estimates can often be obtained from AGQ with a few, e.g. 5–10 nodes. Following Joe (2008) we recommend that the Laplace approximation is used initially; the final model may be estimated accurately with AGQ by increasing the number of nodes until the parameter estimates do not change by any relevant amount.

The Laplace approximation and AGQ are implemented in R-package ordinal (Christensen, 2012) for CLMMs and AGQ is also implemented in the NLMIXED procedure for SAS (Inc., 2008).

4.1. Attenuation effects

A mixed effects model is known as a *conditional model* because the model is formulated for the conditional distribution of the response given the random effects. This means that the parameters of the model apply at the level of subjects and not at the population level, so these parameters are known as *subject-specific* parameters.

Models like the quasi-CLM and DM model are known as marginal models since these models are formulated for the marginal distribution of the response $E_B[Y]$. Usually in such models the correlation structure is treated as a nuisance and only needed to obtain inference for the mean structure. Since the marginal distribution is modeled directly, the parameters of these models apply at the population level and are denoted *population-average* parameters (Diggle et al., 2002; Agresti, 2002; Fitzmaurice et al., 2009).

 ISSN: 2102-6238

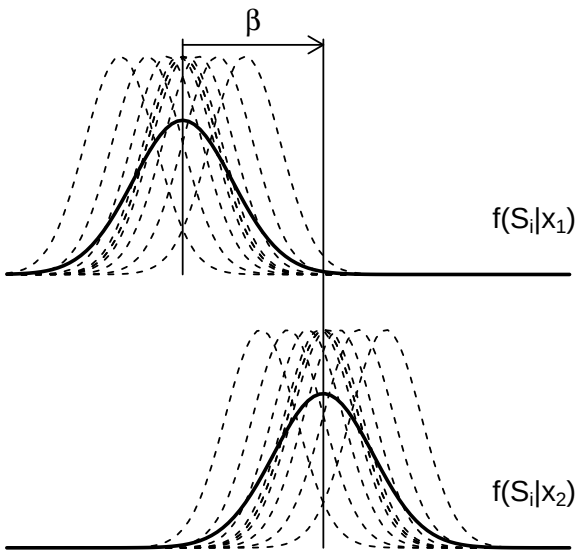

FIGURE 3. *Illustration of the attenuation effect of a CLMM in terms of latent distributions.*

Marginal models do not model individuals explicitly like conditional models, where a subject-distribution is assumed. Conditional models are models for the data-generating mechanism whereas quasi models are not full distributional descriptions.

While it is not possible to obtain subject-specific interpretations from a marginal model, it is possible to obtain population-average interpretations from a conditional model because a particular conditional model implies a marginal model (Zeger et al., 1988; Zeger and Liang, 1992). Marginal predictions and population-average parameters can therefore be obtained in two general ways: 1) by modeling the marginal distribution directly or 2) by obtaining the marginal predictions and parameters implied by a conditional model. Often these two population-average parameter sets are of similar magnitude and usually lead to the same inference. Consequently, conditional cumulative link mixed models constitute a richer framework than the marginal models.

In normal linear mixed models subject-specific and population-average parameters coincide, but in cumulative link mixed models, and generalized linear mixed models in general, the population-average parameters implied by a conditional model are *attenuated*, i.e. smaller in absolute size, relative to the subject-specific parameters. In the CLMM with a probit link, the expectation over the random effects distribution, i.e. the implied marginal model, can be derived explicitly, for details see appendix A:

$$\begin{aligned} E_B[\gamma_{ijl}] &= E_B[\Phi(\theta_j - \boldsymbol{x}_{il}\boldsymbol{\beta} - b_i)] \\ &= \Phi(\theta_j^{pa} - \boldsymbol{x}_{il}\boldsymbol{\beta}^{pa}) \end{aligned} \tag{12}$$

where $\boldsymbol{\theta}^{pa} = \boldsymbol{\theta}/\sqrt{1+\sigma_b^2}$ and $\boldsymbol{\beta}^{pa} = \boldsymbol{\beta}/\sqrt{1+\sigma_b^2}$ are the population-average parameters implied by the conditional model.

 ISSN: 2102-6238

The attenuation effect is illustrated in Fig. 3. Each of the dashed curves represent the latent distribution, $f(S_i)$ for an individual at two predictor values, x_1 and x_2. At x_2 the latent distributions are shifted an amount, β relative to x_1. The solid curves are the latent distributions at the population level which are averaged over individuals. Due to the variation among individuals, the population-average distributions have higher variance than the subject-specific distributions. The relative shift of the curves, that is the size of the shift relative to the spread of the curves, is therefore smaller for the population-average distributions than for the subject-specific distributions.

If all individuals assess a single sample, the indexes i and l coincide (cf. eq. (10) and (12)) and the variance components, σ_b^2 and σ^2 are completely confounded. Thus, if there is heterogeneity among individuals, the estimate of $\boldsymbol{\beta}$ from the non-replicated design (or if individual heterogeneity is not accounted for in a replicated design), $\boldsymbol{\beta}^{pa}$ is attenuated, i.e. too small in absolute size. While the standard tests are valid for non-replicated designs even if there is variation among individuals, the parameter estimates are not consistent and too small in absolute size.

4.2. Tests in marginal and conditional models

In marginal models, inter-individual variation will always translate into over-dispersion, inflation of standard errors and therefore more conservative tests of e.g. product differences. In conditional models this is not always the case. Not only can the naive test; the test ignoring replications all together, be more appropriate than the test in a marginal model with inflated standard errors, the test in a conditional model can also be even more powerful than the naive test. In some cases the naive test will even be unreasonable and a more appropriate test is provided by the conditional model.

This may happen in randomized block settings, that is, in situations with crossed factors, as in example 1, where each consumer evaluated both concordant and discordant product pairs. The randomized block setting is the most common structure for consumer preference studies — although typically with only one evaluation for each combination of product and consumer — or even less in incomplete settings. In the example here, there are replications on top of the randomized blocks, but this is not of key importance for the point to be made here. As opposed to this we have the purely nested ("completely randomized") situation illustrated in example 3, with a grouping of the consumers as the effect of interest — again with additional replications within consumers on top of this.

In the randomized block settings, the proper test for product/treatment differences does not include the block (main) effect — it is removed from the error — this is the main idea of making a blocked experiment. In normal linear models with complete data for an unreplicated randomized block experiment, the sums of squares (SS) decompose into: SS(total)=SS(block)+SS(treat)+SS(error), and the treatment effect is tested against mean square for error. Ignoring blocks in this setting would lead to an error term based on SS(block)+SS(error) rather than SS(error) which in turn leads to a conservative test. Only in situations with either a very weak block effect or a very high number of products relative to the number of blocks, this is not a major problem, but indeed for the typical consumer experiment the pooling of the block effect into the error will grossly affect the analysis.

In analyses of binomial and ordinal data the tests cannot be expressed exactly in terms of mean squares like this, but only approximately so. Consider for example that for binomial data

 ISSN: 2102-6238

with observations away from the extremes and high enough binomial denominator, a normal approximation analysis would give more or less the same results as a standard logistic regression analysis. Clearly, even though the data come in a binomial or ordinal form, it will generally be inadequate to pool the (random) block effect into error, which is exactly what a marginal analysis corresponds to. To summarize, the simple over dispersion approach entailed by marginal models is not well suited to handle random effect models other than purely nested ones. This point is not commonly realized nor even discussed in the literature. A clear advantage of mixed effects models is that they lead to correct tests irrespective of experimental design and effect sizes.

4.3. Example 4

In this example we will revisit the paired degree-of-difference test from example 1 in section 2.4 where 25 subjects each assessed four concordant and four discordant product pairs. A Stuart-Maxwell test adjusted for over-dispersion was suggested in Bi (2002). This test gave $\tilde{X}^2_P = 3.85$, df = 2, $p = 0.146$, which is more conservative than the tests that assumed independent observations. Similarly, for a Wald test in a quasi-CLM we find $\hat{\phi}_G = 1.88$, so $W^* = 1.58$, df = 1, $p = 0.114$ with essentially the same conclusion. A cumulative link mixed model that allows for subject-specific effects reads

$$\gamma_{ijk} = \Phi(\theta_j - p_k - b_i) \quad j = 1,\ldots,3 \quad k = 1,2 \quad i = 1,\ldots,100 \tag{13}$$

Observe that product and subject factors are crossed in the sense of section 4.2, so we can expect the test of product differences to be as strong, or perhaps stronger, in the mixed effects model in comparison with the naive test. The likelihood ratio test of p_k in model (13) is $LR = 5.84$, df = 1, $p = 0.016$, which provides strong evidence of a product difference. Not only is the test stronger than the adjusted Stuart-Maxwell test and the Wald test from a quasi-CLM, it is also stronger than the naive tests reported in example 1 for the same data where individual differences were ignored.

4.4. Tests of random effects terms in cumulative link mixed models

Likelihood ratio tests can be used to test fixed-effects model terms in the same way for cumulative link mixed models as in cumulative link models — tests of random effect terms is a bit more complicated. A likelihood ratio test of a random-effects term is a test of the following hypotheses for the variance parameter:

$$H_0 : \ \sigma_b = 0 \ \text{ versus } \ H_1 : \ \sigma_b > 0 \, . \tag{14}$$

Observe that the test is one-sided, since the random effects standard deviation is non-negative. The usual asymptotic theory for the likelihood ratio statistic, *LR* dictates that the *LR* asymptotically follows a χ^2_1-distribution with one degree of freedom. However, since the σ_b is on the boundary of the parameter space, the usual asymptotic theory does not hold. Following Self and Liang (1987); Stram and Lee (1994) the *LR* more closely follows an equal mixture of χ^2-distributions with zero degrees of freedom (a point mass distribution) and one degree of freedom. The *p*-value from this test can be obtained by halving the *p*-value from the test assuming $LR \sim \chi^2_1$. This adjusted test can be motivated by the following: for a single parameter, we can consider the

 ISSN: 2102-6238

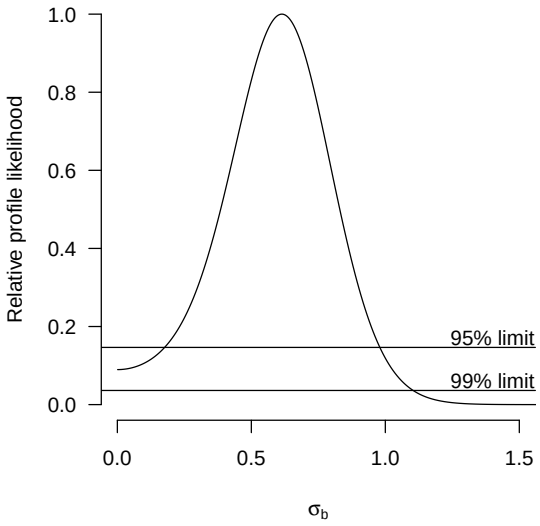

FIGURE 4. *Profile likelihood of σ_b in model (13) for the paired degree-of-difference example.*

likelihood root statistic, $r = \text{sign}(\hat{\sigma}_b - \sigma_0)\sqrt{LR}$; the signed square root of the likelihood ratio statistic (Brazzale et al., 2007; Pawitan, 2001), which under the usual asymptotic theory follows a standard normal distribution. Here, σ_0 is the value of σ_b under the null hypothesis and $\hat{\sigma}_b$ is the maximum likelihood estimate of σ_b. The p-value from the one-sided test of the hypotheses in (14) can be computed as $p = 1 - \Phi(r)$ and is exactly the p-value from the adjusted likelihood ratio test.

Wald tests of the variance parameter can also be constructed, but since the profile log-likelihood function is only approximately quadratic when $\hat{\sigma}_b$ is not small and well defined, such tests cannot be recommended (Pawitan, 2000; Boyles, 2008). Confidence intervals for parameters should preferably be constructed from profile likelihood functions rather than from inverted Wald tests as is for instance implemented in the R-package ordinal (Christensen, 2012).

4.5. Example 5

In this example we continue the analysis of the paired degree-of-difference test from example 1 in section 2.4 and illustrate how inference for the assessor population can be conducted.

The estimated location difference between the two products is 0.404 with standard error 0.168 and the random-effects standard deviation is $\hat{\sigma}_b = 0.614$. Observe that estimate and standard error of the location parameter are larger as expected. The thresholds estimates are $\hat{\boldsymbol{\theta}} = (-0.073, 0.939)$. The relative profile likelihood for σ_b in Fig. 4 displays the evidence in the data about this parameter. The 99% confidence interval includes zero while the 95% confidence interval does not. While a random-effects spread of zero has some support, it is not likely to be considerably larger than one. The one-sided hypotheses in (14) yields $p = 0.014$, but the significance of σ_b was already visible from the profile likelihood in Fig. 4.

The variance parameter can be interpreted as the variation in the subjects' use of the response

 ISSN: 2102-6238

scale, i.e. the variation in the thresholds among the subjects. Roughly 95% of the population will be within $\pm 2\sigma_b = 1.23$, so the shift of the thresholds will span roughly 1.23 units among the population. In comparison, the distance between the thresholds is 1.01 and the product effect is 0.404, so the variation in the population is considerable compared to the distance between the thresholds and the product effect.

4.6. Example 6

In this section we continue the analysis of the consumer liking example in section 3.2 where we found very weak evidence of a difference in liking ($p = 0.119$) in a Wald test adjusting for over-dispersion. A cumulative probit mixed model for these data reads

$$\gamma_{ijk} = \Phi\left(\frac{\theta_j - c_k - b_{i(k)}}{g_k}\right) \quad i = 1,\ldots,n_k \quad j = 1,\ldots,5 \quad k = 1,2\,, \tag{15}$$

where the parenthesized index on b indicates that subjects are nested within cities. The joint LR test of c_k and g_k in model (15) gives $p = 0.35$ which is close to the result by Ennis and Bi (1999) ($p = 0.38$) who took an overdispersion approach. The cumulative probit mixed model confirms that there is no evidence of a difference in liking between the two cities. In this case subjects are nested in cities and the naive test is liberal compared to the tests that take account of replications in line with the discussion in section 4.2. Further, tests from the conditional and marginal models lead to equivalent conclusions. There is, however, a considerable variation among consumers in their perception of the liking scale. The maximum likelihood estimate of σ_b is 0.944 in model (15). The normalized profile likelihood in Fig. 5 confirms that the spread is well-determined. The likelihood root statistic from the one-sided test of σ_b is 12.84 corresponding to a p-value of essentially zero (around $5 \cdot 10^{-38}$).

TABLE 3. *Maximum likelihood parameter estimates (standard errors) in models for the city preference data in Table 2.*

Parameter	Model (9) CLM	Model (9) quasi-CLM	Model (15) CLMM
c_2	–0.183(0.084)	–0.183(0.111)	–0.231(0.208)
g_2	–0.185(0.088)	–0.185(0.115)	–0.154(0.098)
σ_b			0.944
θ	–1.93, –1.28, –0.98, 0.056	–1.93, –1.28, –0.98, 0.056	–3.00, –1.86, –1.34, 0.14
log-lik.	–714.15	—	–658.73

The cumulative link model (CLM), the cumulative link model with over-dispersion adjusted standard errors (quasi-CLM) and the cumulative link mixed model (CLMM) are summarized in Table 3. The parameter estimates for the CLM and quasi-CLM are identical and only the standard errors differ reflecting the adjustment for over-dispersion in the quasi-model. The estimated location and threshold parameters are larger in absolute measures for the CLMM in line with the discussion in section 4.1. Also observe that the standard error of the location parameter is larger in the CLMM than in the CLM. In the quasi-model the standard errors are inflated by the same amount while only the standard error of the location parameter is appreciably bigger in the CLMM.

The parameter estimates for the CLMs have population-average interpretations, i.e. they correspond to the effects that we see at the population level. The parameter estimates in the

 ISSN: 2102-6238

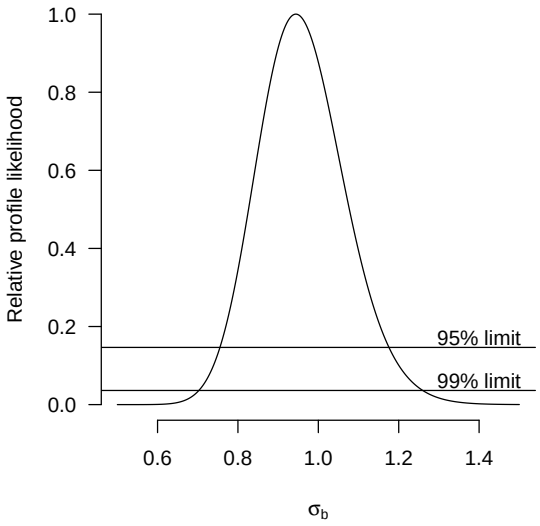

FIGURE 5. *Profile likelihood for σ_b in model (15) for the consumer liking example. The 95% and 99% confidence intervals are indicated by horizontal lines and given by (0.75, 1.18) and (0.70, 1.26) respectively.*

TABLE 4. *The probabilities of rating a product in the five categories for the city preference data in Table 2.*

Segment	dislike very much	2	3	4	like very much
Sample	0.022	0.075	0.069	0.399	0.434
Population-average[a]	0.015	0.079	0.083	0.398	0.424
Subject-specific	0.001	0.029	0.063	0.515	0.392
5% percentile subject	0.077	0.344	0.219	0.335	0.025
95% percentile subject	<0.001	<0.001	0.001	0.078	0.920

[a]: Population-average predictions from CLM and CLMM models coincide.

CLMM have subject-specific interpretations. This is not particularly important for the location and scale differences in this example since these effects are small and insignificant, but it makes a difference in the interpretation of the fitted probabilities. For simplicity of exposition we ignore location and scale differences and consider a CLMM only accounting for subject differences, $\gamma_{ij} = \Phi(\theta_j - b_i)$. The threshold estimates are (–3.11, –1.88, –1.32, 0.27), and the random effects spread estimate is $\hat{\sigma}_b = 1.02$. The probabilities that ratings fall in each of the five categories are presented in Table 4. The first line presents the raw sample proportions. The fitted probabilities from a CLM with no predictors, $\gamma_j = \Phi(\theta_j)$, were identical to the population-average predictions from the CLMM only accounting for subject differences to three digits. These are presented in the second line and are seen to correspond very well to the raw sample proportions. The third line are the fitted probabilities for an average subject, i.e. with $b_i = 0$, which is distinctly different from the probabilities at the population level. From the sample or population estimates we might be tempted to conclude that an average individual would have the highest probability of responding in the "like very much" category because the highest probability is associated with this category, but this is not correct. This kind of subject-specific interpretation should be based on the conditional model predictions presented in line three of Table 4. From this line we see that an average subject is most likely to respond in the fourth category and not the fifth. To illustrate

 ISSN: 2102-6238

the variation between subjects, the ratings of the 5% and 95% percentiles in the distribution of subjects has been included as well reflecting how subjects that are relatively extreme would tend to rate samples. People that like the products the least primarily use the middle categories, while virtually no one would primarily use the "dislike very much" category. On the other hand people that like the products the most almost exclusively rate the products in the "like very much" category.

5. Discussion

In this paper we have shown how cumulative link models can be used for sensory tests on categorical ratings data. We have described how cumulative link models relates to standard omnibus χ^2 tests and how cumulative link models often lead to stronger tests of association because the ordinal nature of ratings data can be utilized. We have suggested two extensions of cumulative link models for replicated data and compared these approaches to the Dirichlet-Multinomial model suggested in Ennis and Bi (1999). Our first suggestion is a quasi-likelihood cumulative link model which leads to Wald tests adjusted for over-dispersion, and our second suggestion is a cumulative link mixed model that explicitly models the population of subjects. When the factor of interest is crossed with the subject factor, marginal models in adjusting for over-dispersion can lead to tests that are weaker and more conservative than naive tests while more correct tests like those of (conditional) mixed models are actually stronger than naive tests. So while approaches adjusting for over-dispersion are not always appropriate, mixed (conditional) models lead to appropriately sized tests irrespective of experimental design. The mixed model can also provide insight into how subjects use the rating scale and can provide subject-specific as well as population-average interpretations. All models discussed in this paper can be fitted with the authors' freely available R-package ordinal (Christensen, 2012).

It was shown in Brockhoff and Christensen (2010) how several common discrimination protocols (m-AFC, duo-trio, triangle and A-not A) can be identified as generalized linear models. This makes it possible to adjust analyses for the effects of e.g. gender differences or varying concentrations of an additive. In this way sensory discrimination and preference protocols are combined with statistical models that enhance the models with a general regression framework. In the same line of work it was shown in Christensen et al. (2011) how the identification of the Thurstonian model for the A-not A with sureness protocol as a cumulative link model with a probit link could allow the analysis of such data to take account of explanatory variables describing the assessors/consumers or the experimental conditions. In this paper we have shown how cumulative link mixed models accommodate replications via random effects. Cumulative link mixed models also extend naturally with a general regression framework and makes it possible to model and control for the effect of explanatory variables — these extensions are also supported by the ordinal package (Christensen, 2012).

In more complicated settings, e.g. in larger consumer preference studies including for instance many consumers, many products and possibly many sessions, it may be of interest to include two or more cross-classified factors as random terms in the model. Cumulative link mixed models with cross-classified random terms can be fitted with the Laplace approximation in ordinal package, while Gaussian quadrature methods are not available for such model structures.

One of the examples considered a degree-of-difference rating experiment. This protocol is

 ISSN: 2102-6238

an extension of the same-different protocol to a rating scale, and while a Thurstonian model has been derived for the same-different protocol, see Macmillan et al. (1977) and Christensen and Brockhoff (2009) for further details, we are not aware of derivations of the Thurstonian model for the degree-of-difference protocol, see, however Irwin et al. (1993) for a discussion. The cumulative link model is not a Thurstonian model for degree-of-difference ratings data per say.

Cumulative link models were also considered for analysis of data from the A-not A with sureness protocol in Christensen et al. (2011). Replicated A-not A with sureness data are also replicated ordinal data, and the methods proposed in this paper can also be used to handle the issue of replications in this situation.

Appendix A: Marginal parameters in a CLMM with a probit link

Following Ten Have et al. (1996) taking the expectation with respect to the distribution of B gives

$$\begin{aligned} E_B[\gamma_{ijl}] &= E_B[\Phi(\theta_j - \boldsymbol{x}_{il}\boldsymbol{\beta} - b_i)] \\ &= E_B[P(Z \leq \theta_j - \boldsymbol{x}_{il}\boldsymbol{\beta} - b_i)] \\ &= P\left(Z \leq \frac{\theta_j - \boldsymbol{x}_{il}\boldsymbol{\beta}}{\sqrt{1+\sigma_b^2}}\right) \\ &= \Phi\left(\frac{\theta_j - \boldsymbol{x}_{il}\boldsymbol{\beta}}{\sqrt{1+\sigma_b^2}}\right) \\ &= \Phi(\theta_j^m - \boldsymbol{x}_{il}\boldsymbol{\beta}^m) \end{aligned}$$

References

Agresti, A. (1999). Modelling ordered categorical data: Recent advances and future challanges. *Statistics in Medicine*, 18:2191–2207.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, second edition.

Agresti, A. and Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review*, 69:345–371.

Bi, J. (2002). Statistical models for the degree of difference test. *Food Quality and Preference*, 13:13–37.

Bi, J. (2006). *Sensory Discrimination Tests and Measurements—Statistical Principles, Procedures and Tables*. Blackwell Publishing.

Boyles, R. A. (2008). The role of likelihood in interval estimation. *The American Statistician*, 62(1):22–26.

Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics—case studies in small-sample statistics*. Cambridge University Press.

Brockhoff, P. B. (2003). The statistical power of replications in difference tests. *Food Quality and Preference*, 14:405–417.

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

Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419):651–658.

Christensen, R. H. B. (2012). ordinal—regression models for ordinal data. R package version 2012.09-12 http://www.cran.r-project.org/package=ordinal/.

Christensen, R. H. B. and Brockhoff, P. B. (2009). Estimation and Inference in the Same Different Test. *Food Quality and Preference*, 20:514–524.

 ISSN: 2102-6238

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

Collett, D. (2002). *Modelling binary data*. London: Chapman & Hall/CRC, 2nd edition.

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in medicine*, 14:1191–1203.

Cressie, N. and Read, R. C. (1989). Pearson's X^2 and the loglikelihood ratio statistic G^2: A comparative review. *International Statistical Review*, 57(1):19–43.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford university Press, 2nd edition.

Ennis, D. M. and Bi, J. (1998). The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies*, 13:389–412.

Ennis, D. M. and Bi, J. (1999). The Dirichlet-multinomial model: Accounting for inter-trial variation in replicated ratings. *Journal of Sensory Studies*, 14:321–345.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer series in statistics. Springer-Verlag New York, Inc., second edition.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molederghs, G. (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC.

Genter, F. C. and Farewell, V. T. (1985). Goodness-of-link testing in ordinal regression models. *The Canadian Journal of Statistics*, 13(1):37–44.

Greene, W. H. and Hensher, D. A. (2010). *Modeling Ordered Choices: A Primer*. Cambridge University Press.

Inc., S. I. (2008). *SAS/STATÂ® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

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

Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Comput. Stat. Data Anal.*, 52(12):5066–5074.

Lawless, H. T. and Heymann, H. (1998). *Sensory evaluation of food*. Chapman and Hall, London.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629.

Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, 84(5):452–471.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42:109–142.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, second edition.

Nair, V. N. (1986). Testing in Industrial Experiments With Ordered Categorical Data. *Technometrics*, 28(4):283–311.

Pawitan, Y. (2000). A reminder of the fallability of the Wald statistic: Likelihood explanation. *The American Statistician*, 54(1):54–56.

Pawitan, Y. (2001). *In All Likelihood—Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the nonlinear mixed-effects model. *Jounal of Computational and Graphical Statistics*, 4(1):12–35.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rayner, J. C. W. and Best, D. J. (2001). *A Contingency Table Approach to Nonparametric Testing*. Chapman & Hall/CRC.

Rayner, J. C. W., Best, D. J., Brockhoff, P. B., and Rayner, G. D. (2005). *Nonparametrics for Sensory Science, A More Informative Approach*. Blackwell Publishing.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. Chapman & Hall/CRC.

Stram, D. O. and Lee, J. W. (1994). Variance Component Testing in the Longitudinal Mixed Effects Model. *Biometrics*, 50:1171–1177.

Ten Have, T. R., Landis, J. R., and Hartzel, J. (1996). Population-averaged and cluster-specific models for clustered ordinal response data. *Statistics in Medicine*, 15:2573–2588.

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

Thurstone, L. L. (1927b). Psychophysical analysis. *American journal of Psychology*, 38:368–389.

Thurstone, L. L. (1927c). Three psychological laws. *Psychological Review*, 34:424–432.

 ISSN: 2102-6238

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wasserman, L. (2004). *All of statistics: A concise cource in statistical inference*. Springer.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the gauss-newton method. *Biometrika*, 61:439–447.

Zeger, S. L. and Liang, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11:1825–1839.

Zeger, S. L., Liang, K.-Y., and Albert, P. A. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44:1049–1060.

 ISSN: 2102-6238