

MODELING HETEROGENEITY IN RANKED RESPONSES BY NON-PARAMETRIC MAXIMUM LIKELIHOOD: HOW DO EUROPEANS GET THEIR SCIENTIFIC KNOWLEDGE?

BY BRIAN FRANCIS^{*}, REGINA DITTRICH[†] AND REINHOLD HATZINGER[†]

Lancaster University^{} and Vienna University of Economics and Business[†]*

This paper is motivated by a Eurobarometer survey on science knowledge. As part of the survey, respondents were asked to rank sources of science information in order of importance. The official statistical analysis of these data however failed to use the complete ranking information. We instead propose a method which treats ranked data as a set of paired comparisons which places the problem in the standard framework of generalized linear models and also allows respondent covariates to be incorporated.

An extension is proposed to allow for heterogeneity in the ranked responses. The resulting model uses a nonparametric formulation of the random effects structure, fitted using the EM algorithm. Each mass point is multi-valued, with a parameter for each item. The resultant model is equivalent to a covariate latent class model, where the latent class profiles are provided by the mass point components and the covariates act on the class profiles. This provides an alternative interpretation of the fitted model. The approach is also suitable for paired comparison data.

1. Introduction. Ranked data commonly arise in many substantive areas such as psychology, social research and marketing research when the interest is focused on the relative ordering of various items, options, stimuli or objects. A typical aim of such studies is to estimate the mean or average ordering of a set of items, and to investigate how this ordering changes with respondent characteristics. This paper focuses on the analysis of a survey question from a special Eurobarometer survey on science knowledge, which asked respondents to rank six sources of science information in order of importance.

Eurobarometer public opinion surveys have been carried out in all member states of the European Union since 1973. Eurobarometer 55.2 was a special survey collected in 2001 and designed to elicit information on European experience and perception of science and technology. 17 countries in

AMS 2000 subject classifications: Primary 62J15, 62J12; secondary 62F07, 62P25

Keywords and phrases: Ranked data, random effects, NPML, paired comparisons, Bradley-Terry model, latent class analysis, mixture of experts, Eurobarometer

Eurobarometer 55.2 May-June 2001 Question 5.		
Here are some sources of information about scientific developments.		
Please rank them from 1 to 6 in terms of their importance to you		
(1 being the most important and 6 the least important)		
a)	Television
b)	Radio
c)	Newspapers and magazines
d)	Scientific magazines
e)	The internet
f)	School/University

FIG 1. *The 'Sources of science information' question*

total were surveyed - with Northern Ireland, Great Britain, East Germany and West Germany being treated as separate countries for the purposes of the survey. Within each country a multi-stage sampling scheme was used. Primary sampling units (PSUs) were randomly selected with probability based on population size after stratification by administrative region and by the degree of urbanization. Within each PSU, a cluster of addresses was sampled, and random route methods were used to select households. Finally, a respondent was selected at random from within each household. Face to face interviewing was used to elicit responses.

Our question of interest in this paper is given in Figure 1. The survey report (Christensen, 2001) describes how this question was analyzed. Only the first two rank positions were examined, and the percentage of times a source was mentioned in either the first or second position was reported. This was presented as given in Table 1:

TABLE 1
Respondents mentioning source of information in first or second position

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>
Television	Radio	Press	Scientific Magazines	Internet	School and University
<i>TV</i>	<i>Radio</i>	<i>Press</i>	<i>SciM</i>	<i>WWW</i>	<i>Edu</i>
60.3%	27.3%	37.0%	20.1%	16.7%	20.3%

This method of analysis however does not use the respondent's last four ranked positions, and also does not distinguish in importance between the first and second ranked position. Thus, information is wasted and other issues such as the influence of covariates and respondent heterogeneity are

not considered.

We proceed by examining current approaches to ranked data in Section 2, before describing our modeling approach in Sections 3 to 5. This approach combines the modeling of ranked data patterns through the Bradley-Terry model. We parameterize the items through a set of worth parameters which sum to 1, and which we allow to depend on covariates. The model also incorporates discrete or nonparametric (mass-point) random effects to account for heterogeneity. This model can also be thought of as a mixture or latent class model on the ranks. Algorithmic and computational issues are discussed in Section 6, and the results of the new analysis on the Eurobarometer question above is discussed in Section 7. The paper finishes with a discussion of the methodology.

2. Existing approaches to ranked data. Three simple approaches to analyzing ranked data are common in the literature. The crudest method is simply to analyze only the first ranked response, but this wastes information by not using the other ranks. Another approach is to assume that the rankings are from a continuous scale, and to analyze mean ranks, perhaps invalidly assuming normality. A third approach, used by sensory perception researchers, uses the non-parametric Friedman two-way analysis of variance. This test however simply examines the null hypothesis that the median ranks for all items are equal, and does not consider any differences in ranking between respondents (Sheskin, 2007). Moreover, if the Friedman test rejects the null hypothesis, no quantitative interpretation such as the odds of preferring one item over another is provided.

All of these simple approaches fail both to consider the underlying psychological mechanism for ranking, and to formulate correct statistical models for this mechanism. In contrast, the approach taken in this paper is statistically more rigorous, and involves modeling the observed ranks by assuming that they are generated through an underlying choice or preference model.

There are also a variety of modeling approaches to ranked data. One common approach assumes that the respondent carries out the ranking by first choosing the most preferred item, and then the next preferred, and so on. This has led to the choice set explosion model of Chapman and Staelin (1982) and the multistage model of Fligner and Verducci (1988). For example, a series of papers by Gormley and Murphy (Gormley and Murphy, 2008a,b) have suggested modeling ranks through the Plackett-Luce and Benter models and have illustrated the methodology using Irish electoral data. However, more generally, the choice set approach has the disadvantage of inconsistency: models which assume instead that respondents first choose

the least preferred, then the next least preferred, and so on lead to different conclusions and estimates of worths.

Other modeling approaches have assumed an underlying distance metric on the ranks - thus [Busse et al. \(2007\)](#) assumed that differences between ranks can be measured through the Kendall distance, which measures the number of adjacent transpositions needed to transform one rank into another. [D'Elia and Piccolo \(2005\)](#) suggested that a two-component mixture of a shifted binomial and a uniform distribution be used to model the rank of an specific item.

In this paper we assume that a ranking of items is produced by the respondent making a set of consistent paired comparison experiments, comparing each item mentally with each of the others, until a consistent ranking is obtained.

[Fligner and Verducci \(1993\)](#) described suitable probability models for ranked data such as the Babington Smith model, where the probability for rankings are defined via parameters for paired comparisons. The usual model for paired comparisons ([Bradley and Terry, 1952](#)) was extended to ranked data by [Mallows \(1957\)](#) (the Mallows-Bradley-Terry model).

[Critchlow and Fligner \(1991, 1993\)](#) showed that the Mallows-Bradley-Terry model is a Generalized linear model (GLM) and extended the model by introducing item-specific variables. We adopt this approach in this paper, extending it by the addition of respondent covariates and random effects structures.

Ranked responses will vary between respondents. While measured covariates can be taken into account ([Dittrich et al., 2000](#); [Francis et al., 2002](#)), there are likely to be other unmeasured or unmeasurable characteristics of the respondents which will also affect the response. This will give rise to heterogeneity in the data which need to be taken into account. One approach is to use a mixing distribution approach. [Lancaster and Quade \(1983\)](#) considered random effects models for paired comparison data and fitted a beta-binomial distribution. [Matthews and Morris \(1995\)](#) later extended the model to involve ties and used a Dirichlet mixing distribution; [Böckenholt \(2001a\)](#) fitted a binomial-Normal distribution.

In this paper, we use a random effects approach, but adopt a discrete non-parametric mass point distribution rather than a continuous mixing distribution. The use of a discrete distribution both avoids the considerable computational complexity of multiple integrals in the continuous case, and also avoids the need to specify a specific distribution which may be inappropriate. Heterogeneity in effect is modeled through the incorporation of a missing latent factor representing group membership. If there are no respon-

dent covariates, then the approach reduces to a latent class model (Formann, 1992). While Böckenholt (2001b), Croon (1989) and Gormley and Murphy (2008a) have considered the use of latent class models for ranked data, they take a choice-based rather than a paired comparison approach.

3. Ranked data and paired comparisons. The ranking of items can be described either by a *rank vector* (which gives the ranks of the items) or by an *order vector* (which gives the items in rank order).

Paired comparisons have much in common with ranking tasks. In a paired comparison task the respondents are asked to choose the preferred item in each pair of items. The number of pairs for a set of J items is given by $\binom{J}{2}$. In general, the observed paired comparison response for two items i and j can be coded as:

$$y_{ij} = \begin{cases} 1 & \text{if item } i \text{ is preferred to item } j, (i \succ j) \\ -1 & \text{if item } j \text{ is preferred to item } i, (j \succ i) \end{cases}$$

It is straightforward to transform a rank order into derived paired comparison data. Suppose the order vector of a respondent on four items is (c, a, b, d) then we know that item c is preferred to item a ; item a is preferred to item b and so on.

However, true paired comparison data and derived paired comparison data from ranks differ in two ways:

- (1) In true paired comparison tasks, respondents might be inconsistent in their preferences, producing an intransitive pattern where the respondent is not choice consistent. In ranking tasks inconsistent response patterns cannot occur.
- (2) The mode of presenting the items is different for the two tasks. In ranking data all items are presented at once while in a paired comparison task all item pairs are presented in turn. Accordingly, different effects concerning the order of the presentation of the items may occur.

4. Modeling ranked data.

4.1. *Modeling a single paired comparison.* The standard approach to modeling paired comparisons is the Bradley-Terry (BT) model (Bradley and Terry, 1952). We define the response in a single paired comparison (ij) to be Y_{ij} . It is assumed that the probability of an item i being preferred to j depends on the non-negative parameters π_i and π_j of the items i and j , defined as follows:

$$P\{Y_{ij} = 1|\pi_i, \pi_j\} = \frac{\pi_i}{\pi_i + \pi_j} \quad \text{and} \quad P\{Y_{ij} = -1|\pi_i, \pi_j\} = \frac{\pi_j}{\pi_i + \pi_j}$$

where we later ensure that the π_i sum to one for identifiability.

Thus,

$$\begin{aligned}
 P\{Y_{ij} = y_{ij} | \pi_i, \pi_j\} &= \left(\frac{\pi_i}{\pi_i + \pi_j} \right)^{\frac{1+y_{ij}}{2}} \left(\frac{\pi_j}{\pi_i + \pi_j} \right)^{\frac{1-y_{ij}}{2}} \\
 (4.1) \qquad \qquad \qquad &= c_{ij} \left(\frac{\sqrt{\pi_i}}{\sqrt{\pi_j}} \right)^{y_{ij}}
 \end{aligned}$$

with $y_{ij} \in \{1, -1\}$ and with a constant $c_{ij}^{-1} = \sqrt{\pi_i/\pi_j} + \sqrt{\pi_j/\pi_i}$ which does not depend on y_{ij} . We now reparameterize π_i as $\lambda_i = \frac{1}{2} \ln \pi_i$ or $\pi_i = \exp(2\lambda_i)$ Equation (4.1) then becomes

$$(4.2) \qquad P\{Y_{ij} = y_{ij} | \lambda_i, \lambda_j\} = c_{ij} \exp(y_{ij}(\lambda_i - \lambda_j))$$

with $c_{ij}^{-1} = \exp(\lambda_i - \lambda_j) - \exp(\lambda_j - \lambda_i)$.

4.2. Response patterns. When transforming ranked data to paired comparison data with J items, we form all possible pairs of items. The number of such pairs is $\binom{J}{2}$ and can be ordered in a standard sequence: $(12), (13), \dots, (1J); (23), (24), \dots, (2J); \dots; ((J-1)J)$. The ranking outcome can therefore be recorded as a paired comparison response pattern vector denoted by $\mathbf{y} = (y_{12}, y_{13}, \dots, y_{J-1,J})$ and consists of a series of 1s and -1s representing the values of the y_{ij} s.

In the case of a true paired comparison task where all possible comparisons are made, the number of all possible response patterns is given by the number of possible outcomes to the power of the number of paired comparisons. If y_{ij} can take only two values, there are $2^{\binom{J}{2}}$ possible response patterns in the space Ω . However, these response patterns also include intransitive patterns which can not be generated from a ranking task. Removing these intransitive patterns, the total number of patterns is considerably reduced to $L = J!$. The space of transitive patterns is denoted by Ω^T . For instance, the intransitive paired comparison pattern $(1 \succ 2, 2 \succ 3, 3 \succ 1)$ has no correspondence with any pattern generated from ranking three items, since ranking patterns are transitive by nature. Incorporation of intransitive patterns in the contingency table would generate structural zeros and neglecting them leads to biased estimates. Therefore the use of a simple BT model, which corresponds to a pattern model including intransitive patterns is not appropriate. Moreover, the dependence introduced by rankings transformed to paired comparisons would not be addressed properly. For instance, assuming independence, and in the simple case of three items, given $Y_{12} = 1, Y_{23} = 1$ the probability of $Y_{13} = 1$ is one, whereas the probability

of $Y_{13} = -1$ is zero. However, modeling the probabilities of whole response patterns and reducing the number of possible patterns to those which are transitive removes these dependencies. We want to emphasize that we only consider complete rankings throughout the paper. It is possible, however, to allow for partial rankings where only a subset of items is ranked (see discussion).

4.3. *Modeling and estimation of transitive response patterns.* The probability for observing a sequence of paired comparisons \mathbf{y} is defined by

$$P(\mathbf{y}) = P(y_{12}, y_{13}, \dots) = \prod_{i < j} P(y_{ij})$$

assuming independence between the comparisons. Using the probabilities for a single paired comparison defined in (4.1) we then get:

$$(4.3) \quad P(\mathbf{y}) = \prod_{i < j} c_{ij} \exp(y_{ij}(\lambda_i - \lambda_j))$$

or equivalently

$$P(\mathbf{y}) = \eta_{\mathbf{y}} \prod_{i < j} c_{ij} \quad \text{with} \quad \eta_{\mathbf{y}} = \exp \sum_{i < j} y_{ij}(\lambda_i - \lambda_j)$$

Parameter estimation is based on multinomial sampling over the transitive paired comparison patterns where it is supposed that each of the N respondents have completely ranked all J items and thus contribute to one of the L transitive response patterns. The probability for observing a certain response pattern \mathbf{y}_ℓ , $\ell = 1, \dots, L$ given J comparisons and transitive relations only is given as

$$(4.4) \quad P(\mathbf{y}_\ell | J, \Omega^T) = \frac{P(\mathbf{y}_\ell)}{\sum_{\ell'=1}^L P(\mathbf{y}_{\ell'})} = \frac{\exp(\eta_\ell) \prod_{i < j} c_{ij}}{\sum_{\ell'} \exp(\eta_{\ell'}) \prod_{i < j} c_{ij}} = \frac{\exp(\eta_\ell)}{\sum_{\ell'} \exp(\eta_{\ell'})},$$

where

$$(4.5) \quad \eta_\ell = \sum_{i < j} y_{ij;\ell}(\lambda_i - \lambda_j)$$

To ease notation $P(\mathbf{y}_\ell | J, \Omega^T)$ is denoted as $P(\mathbf{y}_\ell)$ throughout the paper.

Let n_ℓ be the number of times the response pattern ℓ is observed, then the n_ℓ 's are multinomially distributed where $N = \sum_{\ell} n_\ell$ is the total number of respondents and the probability $P(\mathbf{y}_\ell)$ for a certain response pattern ℓ is given in (4.3).

Thus, the likelihood function is

$$\mathcal{L} = \prod_{\ell} P(\mathbf{y}_{\ell})^{n_{\ell}}.$$

The parameters λ_j can be estimated (using suitable parameter restrictions, e.g., setting the last parameter to zero for identifiability) by using standard software such as the `prefmod` package in R (Hatzinger, 2009). To fit the model a variable containing the counts n_{ℓ} and a specific design matrix \mathbf{X} both need to be set up. The method corresponds to a Poisson log-linear formulation of model (4.4) which is described in detail in Ditttrich et al. (2007), who also describe the more general case when undecided responses can occur.

All parameters in η have interpretation in terms of log odds. Comparing two response patterns ℓ and ℓ' where only one y_{ij} differs, i.e. $y_{ij;\ell} = 1$ and $y_{ij;\ell'} = -1$, the log odds are $\ln(P(\mathbf{y}_{\ell})/P(\mathbf{y}_{\ell'})) = \eta_{\ell} - \eta_{\ell'} = 2(\lambda_i - \lambda_j)$. If the item j is the reference item J , the odds reduce to $\exp(2\lambda_i)$.

Estimates of the worths $\hat{\pi}_j$ are calculated through the expression

$$\pi_j = \frac{\exp(2\lambda_j)}{\sum_j \exp(2\lambda_j)}$$

to ensure that the sum of the worths is equal to 1.

4.4. *Respondent covariates in ranked data.* In most practical applications it is important to determine if the importance of items depend on respondent covariates. This can be viewed as a mixture of experts model. Gormley and Murphy (2008a) give an example analyzing ranked data using a choice-based modeling approach. Initially, we consider categorical covariates only. In this case, each distinct combination of covariates observed will form a covariate set; assume that there are K such sets ($1 < K \leq N$). For example, with two factors AGE (with four levels) and SEX (with two levels), there will be eight covariate sets. To model the effect of the covariates, the $J! = L$ response patterns now become LK response patterns. The number of times the ℓ th response pattern occurs within each covariate set k is denoted by $n_{\ell k}$. The linear predictor η becomes

$$(4.6) \quad \eta_{\ell k} = \sum_{i < j} y_{ij;\ell k} (\lambda_{ik} - \lambda_{jk}).$$

Each λ_{jk} is an interaction effect of the item j and the covariates. Thus, two covariates A and B could potentially lead to the following effects $\lambda_{j.A} +$

$\lambda_{j.B} + \lambda_{j.A.B}$ if an interaction effect on the items between A and B needs to be considered.

With continuous covariates, in general, each respondent will be likely to have his/her own distinct set of covariates, and K will usually be close to N . In the particular example of a single covariate x , the linear predictor of the model generalizes to be of the form

$$\eta_{\ell k} = \sum_{i < j} y_{ij;\ell k} (\lambda_i + x_k \beta_i - \lambda_j - x_k \beta_j)$$

5. The random effects model. While the previous section has allowed for known covariates, there may be other variables which are unmeasured or omitted from the dataset, and these will produce heterogeneity between respondents in the item parameters. One common way to account for such heterogeneity is to introduce random effects for each respondent. These random effects would adjust each item parameter up or down to allow for these missing covariates, and thus we need J random effect components, one for each of the items being ranked.

We now extend the above model to allow for random effects. As before, we work with data aggregated into patterns and covariate sets. For each covariate set and response pattern we need to specify J random effect components $\delta_{j\ell k}$. The linear predictor now becomes

$$(5.1) \quad \eta_{\ell k} = \sum_{i < j} y_{ij;\ell k} (\lambda_{ik} + \delta_{i\ell k} - \lambda_{jk} - \delta_{j\ell k})$$

On the worth scale, the random effects become multiplicative, which will multiply the worths by adjustment factors, shifting the worth for each item up or down in a unique way for each ℓk combination. We set $\delta_{J\ell k}$ to be zero for identifiability, and we define

$$\mathbf{\Delta}_{\ell k} = (\delta_{1\ell k}, \delta_{2\ell k}, \dots, \delta_{J-1;\ell k})$$

a $(J - 1)$ -component random effect vector for each combination of response pattern and covariate pattern.

Integrating over the unknown $(J - 1)$ -component random effects, the likelihood then becomes

$$\mathcal{L} = \prod_{\ell k} \left(\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} P(y_{\ell k} | \mathbf{\Delta}_{\ell k}) g(\mathbf{\Delta}_{\ell k}) d\delta_{1\ell k} d\delta_{2\ell k} \dots d\delta_{J-1;\ell k} \right)^{n_{\ell k}}$$

where $g(\mathbf{\Delta}_{\ell k})$ is the multivariate probability density function or mixing distribution of the random effects vector. For dealing with the multivariate random effect, [Hartzel et al. \(2001\)](#) suggest a number of possible approaches. The first approach is to assume multivariate normality for $g(\cdot)$:

$\Delta_{\ell k} \sim \text{MVN}(0, \Sigma)$, where Σ is an unknown $(J-1) \times (J-1)$ covariance matrix which would be estimated from the data. For example, [Coull and Agresti \(2000\)](#) explored a multivariate binomial logit-normal distribution, where the mixing distribution is multivariate normal.

An alternative method, and one which we explore in this paper, is to adopt a non-parametric solution. This solution replaces the parametric multivariate normal distribution by a series of mass point components with unknown mass or probability, and unknown location. This non-parametric maximum likelihood (NPML) technique ([Aitkin, 1996](#); [Mallet, 1986](#)) has the advantage of being able to identify sub-populations of the respondents with specific response patterns, as well as identifying the effect of respondent covariates on these patterns. The mass-point approach is in fact a mixture model, with the earlier multinomial covariate model being replaced by a mixture of multinomials.

Initially, we suppose that the number of components is known and is set to R . Then we have R mass-point vectors; a typical mass point component r would have unknown mass-point locations

$$\Delta_r = (\delta_{1r}, \delta_{2r}, \dots, \delta_{J-1;r})$$

and unknown component probability q_r . If R is small, this substantially simplifies the problem by replacing a $J-1$ dimensional integral with a sum over R terms.

The likelihood now becomes

$$(5.2) \quad \mathcal{L} = \prod_{\ell k} \left(\sum_{r=1}^R q_r P_{\ell kr}(\mathbf{y}_{\ell k} | \Delta_r) \right)^{n_{\ell k}} \quad \text{where} \quad \sum_{\ell} P_{\ell kr} = 1, \quad \forall k, r$$

The model can be interpreted in two ways. If we consider the discrete mass point components as approximating an underlying multivariate distribution, then we should ignore any interpretation of the mixing structure and interpret the λ_{jk} alone. However, we can also think of the model as representing underlying sub-populations (or latent classes) of the respondents, and we can then interpret the δ_{jr} (which for a specific latent class r gives the extra increase or decrease in item j 's parameter over the reference latent class R).

We determine the number of mass point components by choosing the model which minimizes the Bayesian Information Criterion (BIC) proposed by [Schwarz \(1978\)](#), which provides a penalty on the deviance which is a function of the number of pattern-covariate sets,

$$\text{BIC} = -2 \ln \mathcal{L} + p \ln(LK),$$

where LK represents the number of pattern-covariate combinations and p is the number of parameters in the model.

We need to make clear that the likelihood in (5.2) does not necessarily account for the complex sampling design in the Eurobarometer survey. As the latent classes account for heterogeneity, it is likely that some of the latent classes will reflect clustering and design effects. We return to this point later in the discussion section.

6. Algorithmic and computational issues. The EM algorithm provides a computationally elegant solution to the maximization of the the likelihood given in equation 5.2 (Aitkin, 1996). The use of this algorithm is well known; we give brief details here and provide more detail in the on-line supplement. We start by observing that we can view the problem as a missing data problem, where the latent class membership indicators for each pattern and covariate set are missing. We can write these as $z_{\ell kr}$, with $z_{\ell kr} = 1$ if pattern ℓk belongs to class r , and zero otherwise. The expected values of the z s are defined to be $w_{\ell kr}$ and are the posterior probabilities of class membership for a respondent with pattern ℓ and covariate set k . The E-step of the EM algorithm computes the conditional expectation of the complete log-likelihood (involving the calculation of the w s), whereas the M-step maximizes the multinomial likelihood with respect to the λ s and δ s, given the current expected values of the z s, which can be carried out through an expanded Poisson log-linear model with weights $w_{\ell kr}$. Fitting the multinomial through a Poisson log-linear model necessitates that a set of nuisance parameters be included in the linear predictor; these constrain the marginal totals for each covariate set to be equal to the observed totals.

The $w_{\ell kr}$ can potentially be used to assign respondents to classes. If an respondent belongs to covariate set k and has response pattern ℓ , then we can assign to the class with the highest posterior probability $w_{\ell kr}$ over the r classes.

There are a number of specific problems related to the fitting of latent class models of this kind. The first is that of multiple maxima. The EM algorithm guarantees convergence to a local maximum of the likelihood, but not to a global solution. To minimize this problem, we chose fifty different sets of starting values for each value of R and for each covariate model, and quote the best value of $-2 \ln \mathcal{L}$ and BIC found.

The second problem relates to the well-known slow convergence of the EM algorithm. A relatively tight convergence criterion of 0.001 on the deviance difference was chosen to ensure convergence of parameter estimates.

Additionally, the EM algorithm does not give correct standard errors for

the parameters, as the method assumes that the z s are known rather than estimated. Two solutions are used in this paper. Firstly, it is possible to adopt a hybrid scheme where the EM algorithm is used to obtain convergence, and then a series of Gauss-Newton steps are used to obtain the full Hessian matrix (Aitkin and Aitkin, 1996). A second method which is appropriate where the likelihood is likely to be non-quadratic is to use a procedure described by Aitkin (1994) and Dietz and Böhning (1995) to obtain correct standard errors. This sets the Wald test statistic equal to the likelihood ratio chi-squared statistic obtained by equating one of the parameters in the model to zero. From the Wald-test statistic, the appropriate standard error is obtained for the λ s associated with effect X ,

$$s.e.(\hat{\lambda}_{jX}) = \frac{\hat{\lambda}_{jX}}{\sqrt{2 \ln L(\lambda_{jX} = \hat{\lambda}_{jX}) - 2 \ln L(\lambda_{jX} = 0)}}.$$

It is important that this second procedure is carried out by using as starting values the final estimates of $w_{\ell kr}$ obtained from the final model. This will ensure that the algorithm will not converge to a local maximum with higher deviance.

Both methods have advantages. The first method, while computationally complex, gives asymptotic standard errors for all estimated parameters, provided that good starting values are used for the Gauss-Newton steps. The second method has the advantage of providing a standard error which gives a t-test p-value equivalent to the appropriate likelihood ratio test. However, label switching problems can occur in using the second method especially when setting for example a specific delta parameter to zero.

Finally, for large K , the algorithm will take longer to converge and require more memory, both because of the need to increase the size of the table $[\mathbf{y}_{\ell k}]$ to be analyzed, and the large number of lambda parameters λ_{jk} and nuisance parameters needed to fit the multinomial by means of a Poisson log-linear model. Numerical procedures such as those described in Hatzinger and Francis (2004) can be used to remove the need to estimate the nuisance parameters and to speed convergence.

For this paper, models were fitted using the `pattnpml.fit` function of the R (R Development Core Team, 2009) package `prefmod` (Hatzinger, 2009). The `pattnpml.fit` function is a modification of the `alldist` function in the package `npmlreg` (Einbeck et al., 2007), and has been adapted to allow multiple random effects terms and more flexibility in the choice of start values.

7. Data Analysis. We now apply the above model to the Eurobarometer question. There are 12216 complete responses in the dataset. We choose covariates of **AGE** (4 levels: 15-24, 25-39, 40-54 and 55+) and **SEX** (2 levels: male, female) to illustrate the methodology. There are other important covariates, such as educational level, income and country of origin which have been identified by Christensen (2001) but we exclude these in this illustration to ensure that omitted variables and random effects are needed in the analysis. Of the 720 response patterns, the most popular response is $(TV, Rad, Press, SciM, WWW, Edu)$ with 526 respondents, followed by $(TV, Rad, Press, SciM, Edu, WWW)$ with 507. Only 70 (9.7%) of the response patterns are not used at all by the respondents.

7.1. *Modeling “Sources of science information” data.* Our model fitting strategy was to determine a covariate model using simple fixed effects models (that is, without random effects terms), then fixing the covariates in the model and increasing the number of mass point vectors to allow for the unknown random effects distribution to be approximated by the nonparametric mass point components. We started with the “null” model without covariates (4.5), which estimated a common set of item parameters for all respondents. We then included the respondent covariates **AGE** and **SEX** and examined possible main effect and interaction models. Equation (4.6) reminds us that when we refer to the model **SEX**, we are in fact fitting an interaction term between the *items* $(TV, Rad, Press, SciM, WWW, Edu)$ and **SEX** and specifying 12 interaction parameters in the model: $\lambda_{TV.SEX}$, $\lambda_{Rad.SEX}$, $\lambda_{Press.SEX}$, $\lambda_{SciM.SEX}$, $\lambda_{WWW.SEX}$ and $\lambda_{Edu.SEX}$. Two of these parameters ($\lambda_{Edu.male}$ and $\lambda_{Edu.female}$) are constrained to zero for identifiability. We examined changes in deviance and the Bayesian information criterion BIC (Schwarz, 1978) to compare model fits and to find the best model (that is, the model with the lowest BIC). To allow deviances and BIC values to be compared, we fitted models to the same sized table $[y_{\ell k}]$ - with eight covariate sets, all model fits included eight nuisance parameters (the **AGE** by **SEX** interaction).

As can be seen in Table 2, the main effects model **AGE+SEX** has the lowest BIC (= 18100) and there is no need for the interaction between **AGE** and **SEX**. In the paired comparison model this means both factors **AGE** and **SEX** have a separate effect on the item parameters and therefore the worths of the items change with **AGE** and **SEX**.

TABLE 2
Fixed effect models

model	deviance	no. of parameters	BIC
null	21293	13	21406
AGE	18078	28	18321
SEX	21041	18	21197
AGE+SEX	17815	33	18100
AGE+SEX+AGE:SEX	17790	48	18206

TABLE 3
NPML random effects models with and without covariates

(a) without covariates				(b) with AGE and SEX as covariates			
No. of mass points r	Deviance	No. of parameters	BIC	Deviance	No. of parameters	BIC	Final model
1	21293	13	21406	17815	33	18100	
2	12494	18	12650	10731	38	11060	
3	10252	23	10451	9056	43	9428	
4	9792	28	10035	8836	48	9252	
5	9544	33	9830	8729	53	9187	
6	9387	38	9716	8667	58	9170	✕
7	9302	43	9674	8636	63	9182	
8	9277	48	9693	8623	68	9212	

We can consider two forms of random effects models. We first investigated whether a simple random effects model without covariates provides a better explanation than the fixed effects model. The model without covariates is equivalent to fitting a latent class model to the data. We then fitted random effects models with fixed covariate terms **AGE+SEX**, and tested whether the covariates are still important.

The model with a single mass point component means that all respondents are in one latent class, and corresponds to the null fixed effect model (deviance = 21293). Increasing the number of mass point components (Table 3a), we observed that the BIC steadily decreases with no sign of a minimum being reached. We stopped at eight mass point components, as we were not specifically interested in determining the number needed for the model without covariates. However, we can observe two features. Firstly, through examination of BIC values, the latent class model with two (BIC = 12650) or more components fits substantially better than the covariate model without random effects **AGE+SEX** (BIC = 18100). Secondly, a large number of latent classes will be needed to fully represent omitted covariates

TABLE 4
Parameter estimates for $\lambda_{SciM.AGE}$ for fixed and random effects models: AGE+SEX

AGE	(a) Fixed effects model		(b) Mixture random effects model		
	estimate	standard error	estimate	raw EM standard error	corrected standard error
15-24	0	-	0	-	-
25-39	0.165	0.011	0.169	0.012	0.018
40-54	0.201	0.012	0.198	0.013	0.019
55 +	0.219	0.011	0.208	0.013	0.019

(which in this model also include AGE and SEX).

Can a mixed model provide a way forward, and are the measured covariates still important given the importance of latent class structure? Table 3b shows the results obtained by fitting the random effects model with fixed covariates AGE+SEX. With one mass point component, the model corresponds to the fixed effects AGE+SEX model in Table 2. The minimum BIC is found at $r = 6$ classes; the deviance is substantially less than the deviance for $r = 8$ classes with no covariates. It appears that the fixed effects provide additional explanatory power, and this becomes our final model. Removal of AGE and SEX in turn produces a large significant change in deviance and the covariate model cannot be simplified.

We can interpret the final fitted model in two ways. We can treat the mass point components as approximating an unknown multivariate distribution, and focus attention primarily on the covariates. As an illustration, Table 4 shows the estimates for $\lambda_{SciM.AGE}$ for both the fixed effects model and the final random effects model, with a reference category of school/university (Edu). We can see that as age increases, the preference for scientific magazines compared to school/university as a source of information increases - this is true for both fixed and random effects models, but the effects are attenuated for the random effects model. Other age parameters (not shown) show a relative preference decrease in the use of the internet (WWW), and an increase in TV, newspapers (Press) and scientific magazines compared with school/university. Unadjusted and corrected standard errors (Aitkin, 1994) are given for the random effects model and we can observe that the uncorrected and corrected standard errors are relatively close in this example.

From the estimates of $\lambda_{items.SEX}$ (not shown), we can also conclude that the preference for both scientific magazines and the internet relative to school/university is significantly lower for females than for males.

TABLE 5
Proportions in the six classes

	class 1	class 2	class 3	class 4	class 5	class 6
proportions of patterns	0.3156	0.1289	0.3329	0.0583	0.0984	0.0659
proportions of respondents	0.1808	0.0739	0.2460	0.0716	0.1407	0.2890

It is also possible to proceed by treating the mass point components as latent classes. Table 5 shows the estimated proportions of patterns \hat{q}_r (which are obtained directly from the algorithm) and the estimated proportions of respondents which are weighted averages of the posterior probabilities of pattern membership in each class ($w_{\ell kr}$), weighted by the proportion of respondents in each pattern. Equations (3) and (4) in the online supplement provide further details. Examining the proportions of respondents, we see that Class 6 is the largest class with just under 29% of respondents, followed by class 3 with about 25% and class 1 with just over 18%.

Figure 2 shows the estimated random effect components Δ_r for all items and all classes (apart for the reference item J and class R which are set to zero) including 95% confidence intervals based on the corrected estimated standard errors. The bars (δ_{jr}) are half the log odds ratios comparing the extra effect of item j to the reference item J (education) and for class r related to the reference class R (class 6).

It can be seen for example, that for class 1 the odds for TV and Radio are substantially lower than for Education compared to class 6 (TV: $\exp(-0.84 \cdot 2) = 0.186$, Radio: $\exp(-0.77 \cdot 2) = 0.215$). In class 4 the odds for Press compared to Education are about 1.5 times higher and for WWW 2.1 times higher than in class 6 (Press: $\exp(0.21 \cdot 2)$, WWW: $\exp(0.37 \cdot 2)$).

Figure 3 shows, for males and for females, the plotted worths against age for each of the six sources of information, for two of the six latent classes. We see that the two classes represent different preference patterns in the data. Class 6 represents a large sub-population who prefer to obtain most of their scientific information from non-text and non-scholarly sources. For all age groups and for both males and females, TV is the most preferred source, with radio the second most preferred and increasing in preference with age. Class 1, in contrast, represents a smaller sub-population which prefers academic sources of information over more popular information sources. In this class,

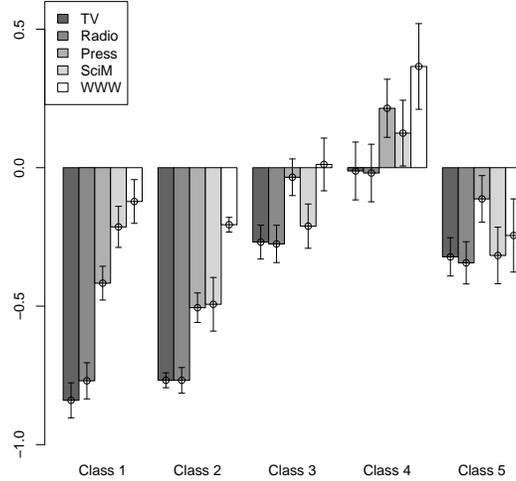


FIG 2. *Parameter estimates for Δ_r and 95% confidence intervals based on corrected standard errors*

for all but the youngest age group, scientific magazines and school/university sources rank in the top two places (with scientific magazines winning out over school/university for males but not for females). For the youngest age group, the school/university followed by the internet are preferred for both males and females. Class 3, the second largest group (not shown), shows a latent class which is similar to class 6 but with a different second preference. TV is still the most preferred source, followed by newspapers and the radio for the three older age groups. For the youngest age group, radio declines in preference and the third preferred source becomes the internet for males and school/university for females. In terms of the other classes which are not displayed, classes 4 and 5 also have TV in first place, but with different orderings of other sources in other places. Class 2 (7%) prefers school/university as the most preferred source of information but with TV in second place.

7.2. Analysis of class membership. It is to be expected that relevant variables not included in the model are absorbed in the latent classes. This relates to variables which are (i) known but for various reasons not accounted

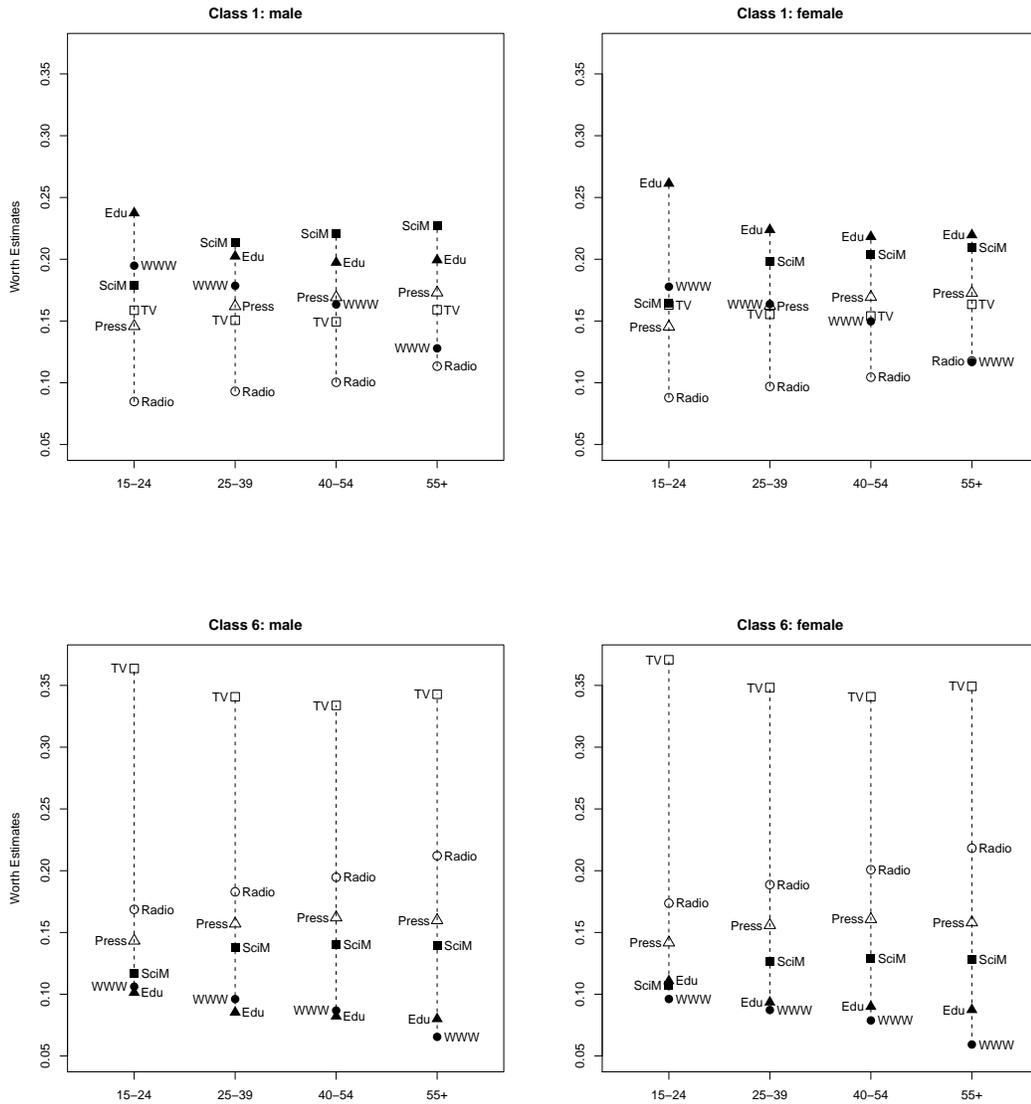


FIG 3. Item worths by age and gender for two extreme latent classes

for (e.g., variables with many categories making computation unfeasible or impossible) and also to (ii) possibly unknown sources of variation. In the Eurobarometer survey, for example, there is a complex five-level clustering design of households within address clusters within PSUs within urbanization and administrative region strata and within countries. While some of these variables are present in the dataset, others are not. In addition, each country has used a different coding scheme for determining degree of urbanization. This means that a full multi-level analysis taking account of all design components is not possible. However, it could be argued that the most important strata are degree of urbanisation and country, and these two levels would account for most variability within the clustered sample. We therefore examine the effect of these two variables below.

To evaluate the effect of known variables a post-hoc analysis may be performed by analyzing their association with the respondents' class memberships. Two approaches are possible which use different definitions of class membership. We illustrate using two covariates not in the model but which are used in the sample design - degree of urbanization and country. For degree of urbanization, we adopt a common three-level categorization which is consistent across countries. We use 15 countries rather than 17 for this investigation, combining East and West Germany (D), and Great Britain and Northern Ireland (GB). The remaining countries are labeled by their international licence plate country code.

The first method uses the posterior probabilities of class memberships to construct the expected number of respondents in each class within each category of the covariate of interest (see equation (4) in the online supplement). We present two mosaic plots ([Hartigan and Kleiner, 1984](#)) which cross-classify the expected class membership with degree of urbanization and with country (displayed in Figure 4).

In examining the degree of urbanization mosaic plot, it can be seen that the proportion of rural residents are underrepresented in class 1 and have a higher proportion in class 6 as opposed to residents of large cities. The country mosaic plot shows much greater variability. Respondents in Italy, for example, are far less likely to belong to latent class 6 and far more likely to belong to latent class 1. In contrast, respondents in Austria and Germany are far more likely to belong to class 6. One explanation for this variability might be the varying quality of TV across countries in broadcasting science information, coupled with a large number of excellent science magazines in Italy.

A second approach, as mentioned in section 6, assigns the respondents (who belong to covariate set k and have response pattern ℓ) directly to

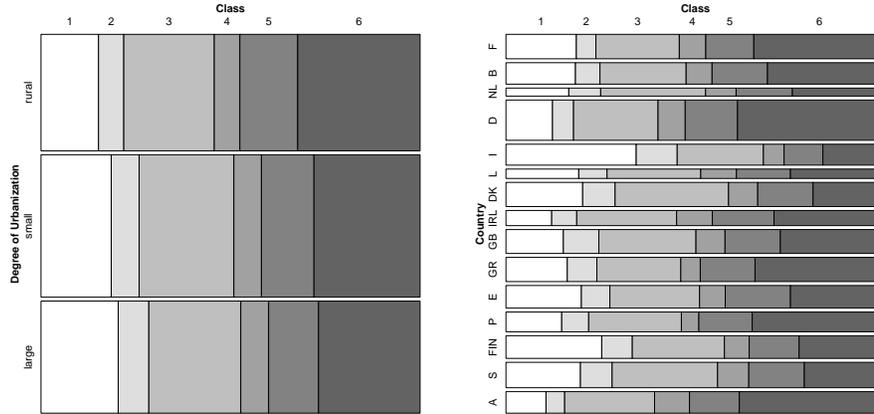


FIG 4. Mosaic plots showing expected class membership and degree of urbanization (left) and country(right)

the class with the highest posterior probability $\max_r(w_{\ell kr})$. Following this procedure, we can obtain a response variable with categories according to the classes and investigate the effects of some variables not included in the model via a multinomial regression model. We then form a cross-classified table of assigned class by country and by degree of urbanization to evaluate possible influences due to part of the multistage sampling design. By fitting a multinomial model we found a strong interaction effect between degree of urbanization and country.

This interaction can be visualized by examining observed log-odds ratios in the constructed table. Figure 5 shows the observed log-odds ratios comparing class 1 to class 6 for the 15 countries both for rural areas and for large cities. We can notice for example that Italy has a positive log-odds ratio for both rural areas and large cities, indicating the relative underrepresentation of class 6 is true both for urban and rural locations. In other countries such as Finland, class 6 is more prevalent in rural areas, and class 1 in large cities.

8. Discussion. Random effects models are often necessary in models for ranked and paired comparison data but the multivariate nature of random effects in these type of models adds complexity. NPML methods of the type described here provide a suitable way forward. The models give greater insight into the nature of subgroups in the dataset, but interpretation can be problematic because of the number of parameters being estimated. We recommend the use of graphical displays on the worth scale.

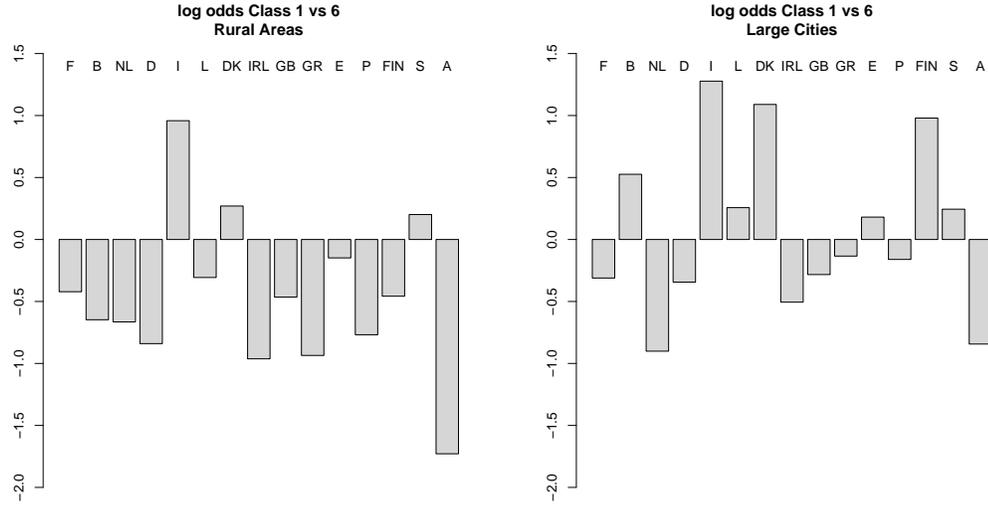


FIG 5. Plot of observed log-odds ratios for class 1 against class 6 for assigned class membership classified by country and degree of urbanization

Diagnostic checks are important for these models. It is important to examine the solution to check both that there are no overly small latent classes, and also that the parameter estimates for each mass point component are sufficiently separate (McLachlan et al., 1999). Posterior probabilities of component membership could also be examined in relation to other covariates not in the model to aid interpretation of the latent classes (Kamakura and Mazzon, 1991).

The basic model described in this paper can be extended in various ways.

- Extensions to models which allow varying coefficients with latent classes is straightforward. This model will allow for different respondent covariate effects within each latent class. These random coefficient models can be fitted by allowing interactions between the latent class group and the covariates, but with the disadvantage of a sizeable increase in the number of model parameters.
- It is possible to extend the model to allow for tied ranks. Such data will lead to an underlying ordinal paired comparison model (Dittrich et al., 2004).
- Item covariates could also be included along the lines suggested by Dittrich et al. (1998).

- The model presented here needs to be extended to allow explicitly for more complex sampling designs and other multi-level structures which may be present in the data. Further research is needed on this topic.
- Finally, incomplete or partial rankings could also be taken account of. This would lead to a paired comparison model which allows for missing comparisons within a response. The basic idea here is to extend the set of response patterns to include patterns where certain comparisons are not available. For partial rankings a composite link approach to this problem has been described in [Dabic and Hatzinger \(2009\)](#), the general case for paired comparisons with missing data is treated in [Dittrich et al. \(2010\)](#). Unfortunately, the number of response patterns may increase dramatically and thus this approach is computationally feasible only for a small number of items.

In conclusion, our approach provides a methodology which allows the modeling of ranked data in many applied areas, allowing covariates to be taken into account and latent classes to be detected. The underlying paired comparison approach provides an attractive alternative to the choice based models dominant in the literature.

Acknowledgements. This research was supported by the ESRC under the National Centre for Research Methods initiative (grant numbers RES-576-25-5020 and RES-576-25-0019). We would like to thank Walter Katzenbeisser for helpful statistical discussions, and the referees and editors for insightful suggestions. Eurobarometer questions are reproduced with the license granted by its author, the European Commission, Directorate-General for Information, Communication, Culture and Audiovisual Media, 200 rue de la Loi, B-1049 Brussels, and by permission of its publishers, the Office for Official Publications of the European Communities, 2 rue Mercier, L-2985 Luxembourg (© European Communities). The dataset was provided by Zentralarchiv für Empirische Sozialforschung, Köln (ZA 3509).

References.

- Aitkin, M. (1994). An EM algorithm for overdispersion in Generalised Linear Models. In Hinde, J., editor, *Proceedings of the 9th International Workshop on Statistical Modelling*, pages 1–8. Exeter University.
- Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6:251–262.
- Aitkin, M. and Aitkin, I. (1996). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*, 6:127–130.
- Böckenholt, U. (2001a). Hierarchical modelling of paired comparison data. *Psychological Methods*, 6(1):49–66.
- Böckenholt, U. (2001b). Mixed-effects analyses of rank ordered data. *Psychometrika*, 66(1):45–62.

- Bradley, R. and Terry, M. (1952). Rank Analysis of Incomplete Block Designs. I. The Method of Paired Comparisons. *Biometrika*, 39:324–345.
- Busse, L., Orbanz, P., and Buhmann, J. (2007). Cluster Analysis of Heterogeneous Rank Data. In *Proceedings of the 24th International Conference on Machine Learning*, pages 113–120, Corvallis. OR.
- Chapman, R. and Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, 19(3):288–301.
- Christensen, T. (2001). Eurobarometer 55.2: Europeans, science and technology. Technical report, European Opinion Research Group, Commission of the European Communities, Brussels.
- Coull, B. and Agresti, A. (2000). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, 56:73–80.
- Critchlow, D. and Fligner, M. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation in GLIM. *Psychometrika*, 56:517–533.
- Critchlow, D. and Fligner, M. (1993). Ranking models with item variables. In Fligner, M. and Verducci, J., editors, *Probability Models and Statistical Analyses for Ranking Data*, pages 1–19, New York. Springer-Verlag Lecture Notes in Statistics 80.
- Croon, M. (1989). Latent class models for the analysis of rankings. In De Soete, G. and Klauer, J.S., editor, *New developments in psychological choice modelling*, pages 99–121, Amsterdam. Elsevier.
- Dabic, M. and Hatzinger, R. (2009). Zielgruppenadäquate Abläufe in Konfigurationssystemen - eine empirische Studie im Automobilmarkt: Das Paarvergleichs-Pattern-Modell für Partial Rankings. In Hatzinger, R., Dittrich, R., and Salzberger, T., editors, *Präferenzanalyse mit R*, pages 119–150. Facultas, Wien.
- D’Elia, A. and Piccolo, D. (2005). A mixture model for preferences data analysis. *Computational Statistics and Data Analysis*, 49:917–934.
- Dietz, E. and Böhning, D. (1995). Statistical inference based on a general model of unobserved heterogeneity. In Seeber, G.U.H., Francis, B.J., Hatzinger, R. and Steckel-Berger, G., editor, *Statistical Modelling: Proceedings of the 10th International Workshop*, pages 75–82, New York. Springer-Verlag Lecture Notes in Statistics 104.
- Dittrich, R., Francis, B., Hatzinger, R., and Katzenbeisser, W. (2007). A paired comparison approach for the analysis of sets of Likert scale responses. *Statistical Modelling*, 7:3–28.
- Dittrich, R., Francis, B., Hatzinger, R., and Katzenbeisser, W. (2010). Missing observations in paired comparison data. *under revision*.
- Dittrich, R., Hatzinger, R., and Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Applied Statistics*, 47:511–525.
- Dittrich, R., Hatzinger, R., and Katzenbeisser, W. (2004). A log-linear approach for modelling ordinal paired comparison data on motives to start a phd programme. *Statistical Modelling*, 4:181–93.
- Dittrich, R., Katzenbeisser, W., and Reisinger, H. (2000). The analysis of rank ordered preference data based on Bradley-Terry Type Models. *OR Spektrum*, 22:117–134.
- Einbeck, J., Darnell, R., and Hinde, J. (2007). *npmlreg: Nonparametric maximum likelihood estimation for random effect models*. R package version 0.43.
- Fligner, M. and Verducci, J. (1988). Multistage Ranking Models. *Journal of the American Statistical Association*, 83:892–901.
- Fligner, M. and Verducci, J. (1993). *Probability Models and Statistical Analyses for Ranking Data*, volume 80. Springer Lecture Notes in Statistics.

- Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87:476–486.
- Francis, B., Dittrich, R., Hatzinger, R., and Penn, R. (2002). Analysing ranks using paired comparison methods: an investigation of value orientation in Europe. *Applied Statistics*, 51(3):319–336.
- Gormley, I. and Murphy, T. (2008a). A mixture of experts model for rank data with applications in election studies. *Annals of Applied Statistics*, 2:1452–1477.
- Gormley, I. and Murphy, T. (2008b). Exploring voting blocs within Irish electorate. *Journal of the American Statistical Association*, 103:1014–1027.
- Hartigan, J. and Kleiner, B. (1984). A mosaic of television ratings. *American Statistician*, 38(1):32–35.
- Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, 1(2):81–102.
- Hatzinger, R. (2009). *prefmod: Utilities to fit paired comparison models for preferences*. R package version 0.8-17.
- Hatzinger, R. and Francis, B. (2004). Fitting paired comparison models in R. Technical Report 3, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien.
- Kamakura, W. and Mazzon, J. (1991). Value segmentation? A model for the measurement of values and value systems. *Journal of Consumer Research*, 18:208–218.
- Lancaster, J. F. and Quade, D. (1983). Random effects in paired-comparison experiments using the Bradley-Terry model. *Biometrics*, 39:245–249.
- Mallet, A. (1986). A maximum likelihood estimation method for random coefficient regression models. *Biometrika*, 73:654–656.
- Mallows, C. (1957). Non-null ranking models: I. *Biometrika*, 44:114–130.
- Matthews, J. and Morris, K. (1995). An application of Bradley-Terry-type models to the measurement of pain. *Applied Statistics*, 44:243–255.
- McLachlan, G., Peel, D., Basford, K., and Adams, P. (1999). The EMMIX software for the fitting of mixtures of Normal and t-components. Technical report, Department of Mathematics, University of Queensland.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.
- Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall, London.

DEPARTMENT OF MATHEMATICS AND STATISTICS
 FYLDE COLLEGE
 LANCASTER UNIVERSITY
 LANCASTER LA1 4YF
 U.K.
 E-MAIL: B.Francis@Lancaster.ac.uk

DEPARTMENT OF STATISTICS AND MATHEMATICS
 VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
 AUGASSE 2-6
 A-1090 WIEN
 AUSTRIA
 E-MAIL: Regina.Dittrich@wu.ac.at
Reinhold.Hatzinger@wu.ac.at