ESRC National Centre for Research Methods

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Latent classes for preference data

Brian Francis
Lancaster University, UK

Regina Dittrich, Reinhold Hatzinger, Patrick Mair
Vienna University of Economics, Austria

ESRC National Centre for Research Methods

NCRM Working Paper Series

8/06

Latent classes for preference data

Brian Francis

Lancaster University, UK

Regina Dittrich, Reinhold Hatzinger, Patrick Mair

Vienna University of Economics

Introduction

Social surveys often contain questions where the response is a ranked set of items.

Becoming a Social Worker: Transition through Training project

68. Please order the following possible aims of social work in order, from 1 to 6 (1 being the most important and 6 the least important)

a) to work towards a more fair and just society .....
b) to provide care for people .....
c) to help people maintain relationships.....
d) to ensure that the law is upheld .....
e) to prevent harm to vulnerable people .....
f) to ensure that people's rights are exercised ..

We wish to model the relationship of the ranked response to other covariates in the dataset. But first , why ask this question?

Background - Sociological aim of research

Is there an association between gender and perceived aims of social work?

Gilligan (1982) proposes that

- women's mode of thinking is 'contextual and narrative' and so are concerned with the needs of others and responsibility for others - *"ethic of care"*
- men are "abstract and formal" in their thinking and so primarily directed to ensuring rights, through rules and self-responsibility - "*ethic of justice*".

DISSENT:

Larrabee(1993) gendered divisions may be too simplistic;
Banks (1995) – dangerous to make gendered distinctions – culture may override gender.

Also developmental psychology approach – moral development through "teachable moments" in professional practice (Kohlberg, 1992) – criticised by Gilligan as "Gender Blind".

The Dataset

Five universities in four countries were involved in the study of first year students:

Arizona State University, USA (69 respondents).
University of British Colombia (Canada) (52 respondents),
Curtin University, Australia (20 respondents),
Lancaster University (UK) with 43 respondents,
University of Western Australia (45 respondents)

211 cases, after removing one case with missing age, and 17 cases with inconsistent ranks (eg 1 2 1 2 1 2 3, 2 2 2 2 2 2).

We use gender, age (two categories <30, >=30) and university (five levels) as covariates.

Items consist of three "ethics of care" items and three "ethics of justice" items

Modelling ranked responses

The ranked responses are easily converted to paired comparison form.
For each comparison between two items *i* and *j*, the individual can respond

i preferred to *j* *j* preferred to *i*

We compare each pair of items. In any comparison, if the 1st of a pair gets the lower score, then we say that the 1st item is preferred. If the 1st in the pair gets the higher score, then the 2nd item is preferred.

RANKS: Suppose the rank order given by an individual is

b e a d c f

Then we know that b is preferred to e b is preferred to a
e is preferred to a e is preferred to d etc.

Every respondent generates fifteen paired comparisons.

We prefer this to the sequential discrete choice formulation, which gives results dependent on choice order.

Modelling a single paired comparison (Bradley- Terry model)

We define a response Y_{ij} in the comparison of item *i* to item *j* as follows:

$$Y_{ij} = \begin{cases} -1 & \text{if } j \text{ is preferred to } i \\ 1 & \text{if } i \text{ is preferred to } j \end{cases}$$

We measure the worths of an item *i* through a set of worth parameters π_i, with $\sum_i \pi_i = 1$ for identifiability. Then:

$$P\{Y_{ij} = y_{ij}\} = \Phi_{ij} \left(\frac{\pi_j}{\pi_i + \pi_j} \right)^{1-y_{ij}} \left(\frac{\pi_i}{\pi_i + \pi_j} \right)^{1+y_{ij}}$$

$$P\{Y_{ij} = y_{ij}\} = \Phi^*_{ij} \left(\frac{\pi_i}{\pi_j} \right)^{y_{ij}}, \qquad y_{ij} \in \{-1,1\}$$

The response pattern vector Y

We assume here that there is no missing rank information–we are comparing all possible pairs.

We now define **y** to be the response pattern vector for all paired comparisons generated from the rank response (Critchlow and Fligner, 1993)

$\mathbf{y} = (y_{12}, y_{13}, \ldots, y_{J-1,J})$. Length $\binom{J}{2}$

Each element can take one of the two values (-1 1), so there are $2^{\binom{J}{2}}$ possible response vectors for true paired comparison data.

However, many of these response patterns are intransitive (A<B, B<C, C<A) and cannot be generated from ranked data.

The number of transitive responses is J! For our data, J=6, giving 720 patterns, which we index by l (l=1...720)

Modelling the response patterns - estimation

For each response pattern l, we have

$$P\{\mathbf{y}_l\} = \prod_{i<j} P\{y_{ijl}\} \quad = \prod_{i<j} \Phi_{ij}\left(\frac{\pi_i}{\pi_j}\right)^{y_{ijl}} = \Phi \prod_{i<j} \left(\frac{\pi_i}{\pi_j}\right)^{y_{ijl}}$$

We convert to log-linear form – can be fitted as a standard Poisson log-linear model:

$$m_l = N \; P(y_l)$$

$$\ln(m_l) = \phi + \sum_{i<j} y_{ijl}(\ln \pi_i - \ln \pi_j) \quad = \phi + \sum_{i<j} y_{ijl}(\lambda_i - \lambda_j)$$

where m_l is the expected value for n_l, the number of times the response pattern l is observed, and $N = \Sigma\, n_l$ is the number of respondents.

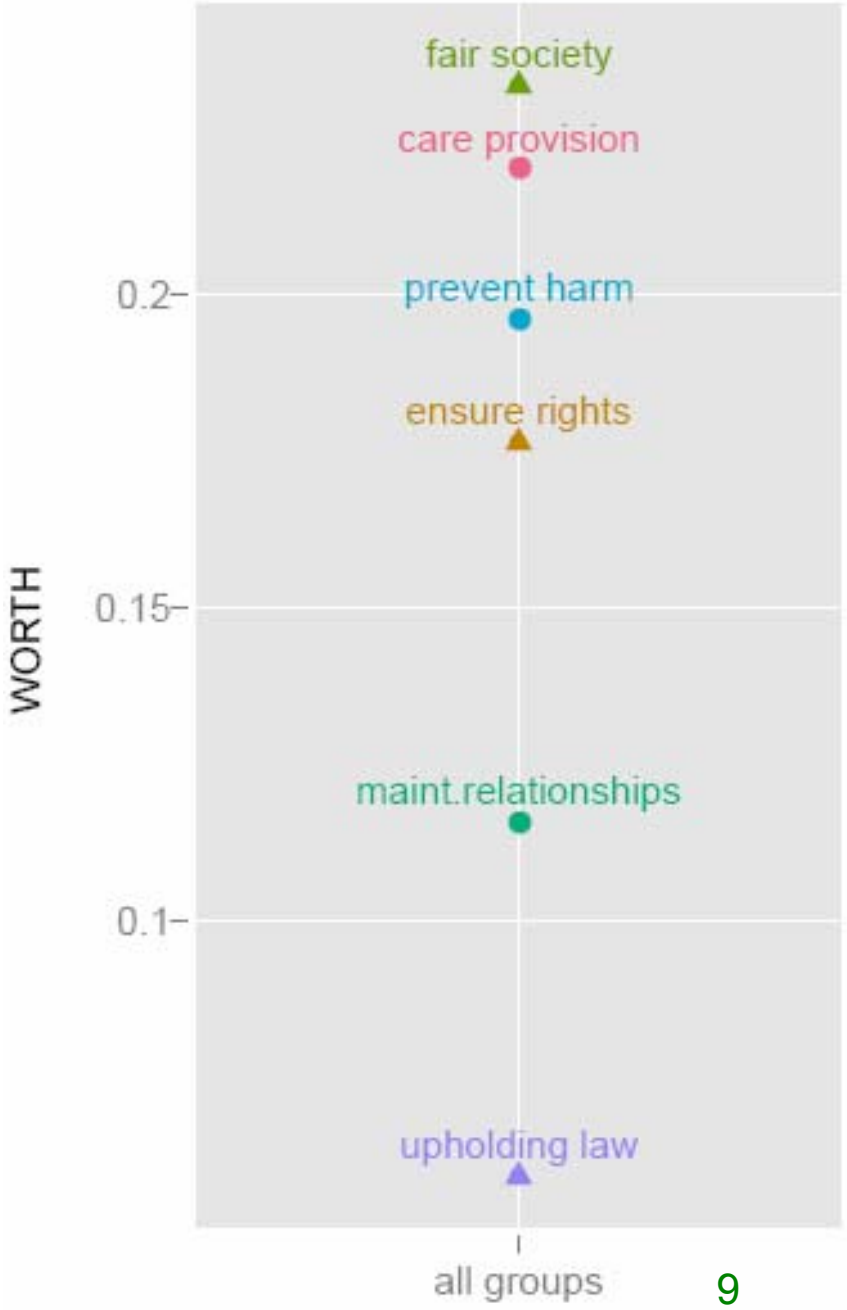
“No covariate” model

We estimate the λ_i - one for each item with $\lambda_J = 0$ for identifiability.

We can then display the worths – the π_i.

We can see that justice aims ▲ and care aims ● are interleaved.

Fair society ▲ and care provision ● are the two items with the highest worths.



Covariates

Assume that values of covariates can be combined into K distinct covariate sets, with $1 < K \leq N$.

Then we expand the data K times, counting the number of times the 720 response patterns occur within each covariate set.

The Poisson log-linear model then becomes:

$$\ln(m_{lk}) = \phi_k + \sum_{i<j} y_{ijlk}(\lambda_{ik} - \lambda_{jk})$$

where m_{lk} is the expected value for n_{lk}, the number of times pattern l is observed in the kth covariate set.
There is now a separate nuisance parameter ϕ_k for each covariate set.

For each covariate level k we estimate (J-1) =5 parameters λ_{jk}, one for each item apart from the reference item

Fixed effect models for student social workers data:

Model (all interacted with items)	Deviance	Δ deviance from model 1	Δ df from model 1	p-value
Model 1. "two-way" Age.sex+age.uni+uni.sex	1019.3			
Age.sex+age.uni	1040.6	21.3	20	0.38
Age.sex +uni.sex	1054.6	35.3	20	0.02
age.uni+uni.sex	1025.0	5.7	5	0.34
Age.uni +sex	1044.2	24.9	25	0.47
Age.uni	1055.1	35.9	30	0.22
Age + uni		78.1	60	0.05

Final model involves age and uni and age.uni interaction (interacted with items) but not sex.
There is little evidence of views of social work being determined by gender at the start of the course.

Random effects in ranked responses

However, there is likely to be heterogeneity in the data (partly due to unobserved covariates) which will need to be modelled.

Need to allow for individual-specific effects due to unmeasured covariates (eg income) and unmeasurable covariates (eg aggressiveness) (degrees of latentness, Crowder).

Individuals are identified only by their response pattern and their covariate set. For each response pattern l and covariate set k, and with J items, we need J-1 random effect components, one for each item. Assume that this adds an effect

$\Delta_{lk} = (\delta_{1li}, \delta_{2lk}, \ldots, \delta_{Jlk})$ onto the item parameters λ_{lk}

δ_{Jlk} defined to be zero for identifiability

$$L = \prod_{lk} \int f(n_{lk} \mid \lambda_k, \phi_k, \Delta_{lk})\, d\Delta_{lk}$$

Random effects in ranked responses

What distribution g(●) do we assume for the Δ_{lk}?

a) Could assume multivariate normality for g.

$\Delta_{lk} \sim$ MVN(0, Σ) where Σ is an unknown J-1 x J-1 covariance matrix
unrealistic and too many parameters to estimate if J large

b) Use a mass point approach

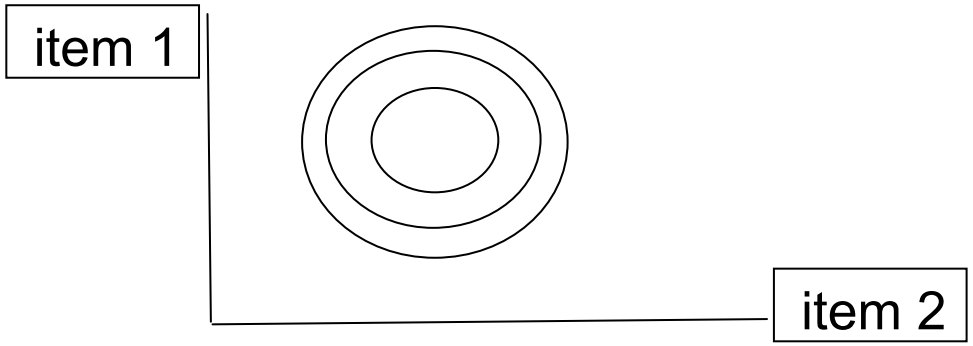
Assume g(●) is a mixture of M mass point vectors Δ_m with probabilities q_m.
Non-parametric maximum likelihood (NPML) estimation of random effects

These are latent class models – each mass point corresponds to a latent class.

Illustration

In two dimensions

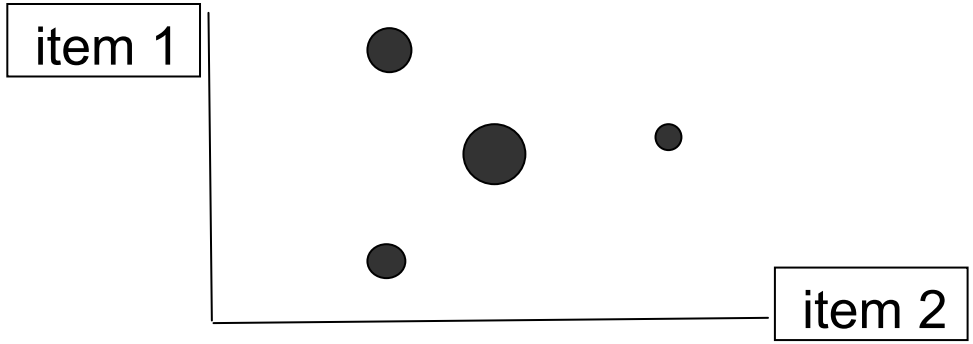
this MVN random effects distribution:



could be replaced by perhaps M=4 mass points at locations

$\Delta_m = (\delta_{1m}, \delta_{2m})$

with probabilities q_m proportional to the size of the dots.



Likelihood is now

$$L = \prod_{lk} \sum_{m=1}^{M} q_m \quad f(n_{lk} \mid \lambda_k, \phi_k, \Delta_m)$$

number of times the pattern l is observed in covariate set k

which is a mixture model, with the Δ_m defining the latent classes

Implementation: Use EM algorithm. Can be fitted as a standard GLM by expanding data M times – details for paired comparison models are trickier.

So we need to expand data MK times to fit covariate models with random effects (latent classes) in standard GLM software.

How many mass points/latent classes to choose?

Needs random start sets to avoid local solutions

Mass points M	Latent class model – no covariates		Latent class model with AGE*UNI as covariates	
	Deviance	# params	Deviance	# params
1 Fixed effects model	999.0	5	918.5	50
2	966.6	10	893.5	55
3	954.8	15	885.5	60
4	952.6	20	883.9	65

Many researchers use BIC or AIC - however in this case table is very sparse (200 persons in 7200 cells) so is unreliable.
Monte Carlo simulation of LR test can also be done but needs care.

We choose a parsimonious model based on theory and examine the AGE by UNI covariate model with 2 mass points / latent classes

Interpretation

Can treat the model either as

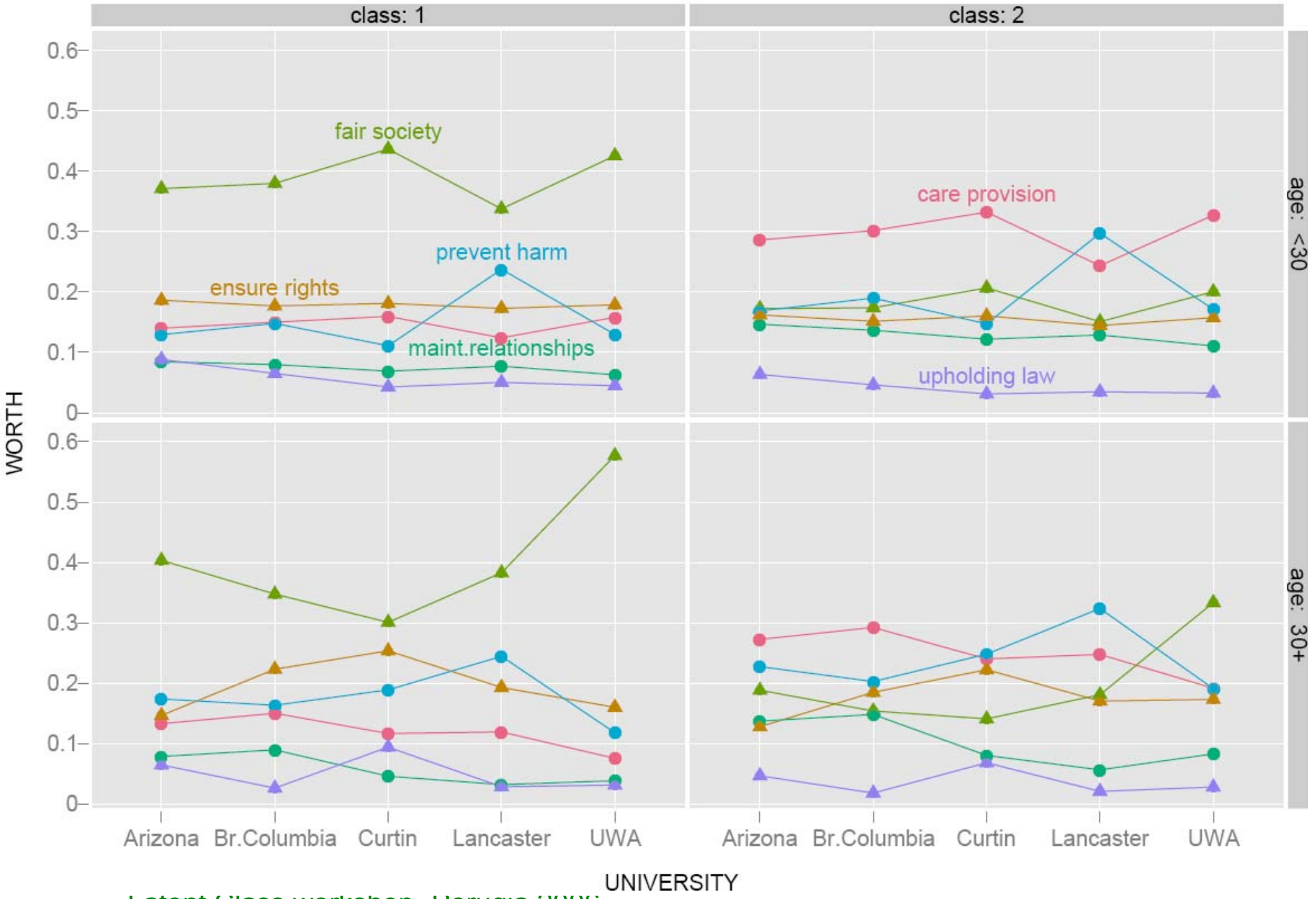
- approximation to underlying unknown continuous Random effects distribution, with interest primarily on measured covariates…
- Or as representing real groups in the data – latent classes have meaning.

EG: Arizona estimates for “Fair and Just society”
(ref category “People’s rights”)

	Fixed effects		2 latent classes	
	Estimate	s.e	Estimate	EM s.e
Age 18-29	0.207	0.158	0.062	0.171
Age 30+	0.484	0.158	0.387	0.171

Age effects are still strong but reduced. Similar effects for university

Can also look at the differential response in the two latent classes:


class: 1
class: 2
age: <30
age: 30+
fair society
prevent harm
ensure rights
maint.relationships
care provision
upholding law
WORTH
0
0.1
0.2
0.3
0.4
0.5
0.6
Arizona
Br.Columbia
Curtin
Lancaster
UWA
UNIVERSITY

Mixture classes- descriptions

58% Class 1 Ethic of justice group Fair society, ensure rights have highest worths in all universities but Lancaster. However, upholding law stays at the bottom of the worth scale.

42% Class 2 Ethic of care group. " Care for people and prevent harm" have highest worths in the older age group for all universities apart from UWA.

Note the consistent placement of "maintain relationships" and "uphold law" in the last two places for all universities and age groups and in both latent classes.

Note also the strong age by university interaction. For some universities, rights-based issues are less important for older students( )for others, rights appears to be more important (Lancaster, UWA)

The gender issue

As we have two latent classes which can broadly be identified as “ethics of justice” and “ethics of care” groups, can we identify any gender effect.

We assign each case to the latent class with the highest posterior probability of class membership.

	Latent class membership	
Gender	Justice	Care
Female	112 (62%)	68
Male	20 (64%)	11

χ^2_1 is 0.0018 , p=0.97. no evidence of gender effect.

Conclusions

- Random effects models are often necessary in models for ranked and preference data but multivariate nature of random effects adds complexity. Latent class methods provide a good way forward.

- Complex interpretation - needs graphical displays

- Provides insight into the gender debate in social work – there appear to be one group who focus on justice and another focusing on care, but equally spread across males and females. Gilligan appears to be wrong in her gendered view of social work. However, do the courses add a gendered component, so that trained social workers are gendered in their views?

- Need to extend model to allow for partial and inconsistent rank responses, but problems with small sample sizes.

Email: info@ncrm.ac.uk
Homepage: www.ncrm.ac.uk