manuscript No.
(will be inserted by the editor)

Markov models of dependence in longitudinal paired comparisons - An application to course design

Alexandra Grand · Regina Dittrich · Brian Francis

Received: date / Accepted: date

Abstract This article suggests a new approach for modelling longitudinal paired comparison data. As individual preferences may change from one time point to another, we propose extending the basic log-linear Bradley-Terry (BT) model by incorporating a Markovian structure with temporal within-comparison dependence parameters and parameters indicating the amount of change of the unknown preference parameters of the objects. We illustrate this approach by analysing a student survey relating to statistics course design with three time points.

Keywords Bradley-Terry model · log-linear model · longitudinal paired comparison data · Markov chain · temporal change of preference parameters

1 Introduction

The method of paired comparisons is a common method for locating objects, items, attitudes etc. on a latent preference continuum (i.e. ordering a set of objects on a scale). Pairing a set of J objects for a paired comparison study, we obtain $\binom{J}{2}$ object pairs (where e.g. the pair of the objects j and k is represented by (jk)).

In this article we are concerned with extending the standard paired comparison method to deal with longitudinal paired comparisons. It is motivated by

A. Grand
Institute for Statistics and Mathematics, WU Vienna, Austria
E-mail: alexandra.grand@gmx.at

R. Dittrich
Institute for Statistics and Mathematics, WU Vienna, Austria
E-mail: regina.dittrich@wu.ac.at

B. Francis
Department of Mathematics and Statistics, Lancaster University, UK
E-mail: B.Francis@Lancaster.ac.uk

a study on students' course design preferences in statistics measured at three time points. The course designs are direct instruction (cd1), self-initiated content preparation (cd2), e-learning with discussion of the solution (cd3) and e-learning (cd4), the four objects of interest. In each of the six paired comparisons at each time point, i.e. (cd1,cd2), (cd1,cd3), (cd2,cd3), (cd1,cd4), (cd2,cd4), (cd3,cd4), students were asked to choose their preferred course design. The example also raises issues of attrition which we account for in our model.

For the analysis of paired comparison data, we refer to the well-known Bradley-Terry model (Bradley and Terry, 1952). The basic BT model has been extended in logistic and in log-linear representation by various authors. Some developments are, for example: the incorporation of object-specific covariates as well as subject covariates (Kousgaard, 1984; Matthews and Morris, 1995; Dittrich et al, 1998; Francis et al, 2002), between-comparison dependencies (Dittrich et al, 2002) and temporal dependencies for longitudinal or panel data (Fahrmeir and Tutz, 1994; Böckenholt, U. and Dillon, W., 1997; Glickman, 1999, 2001; Cattelan et al, 2013; Francis et al, 2014).
Longitudinal paired comparisons are characterized by repeatedly making decisions on the same paired comparisons over time points t, $t = 1, 2, \ldots, T$, by the same individuals (judges). It is assumed that individual preferences may vary from one time point of observation to another time point (that is, that enough time has passed between time points to allow change of opinion).
Fahrmeir and Tutz (1994) defined a non-Gaussian *state-space* model for longitudinal paired comparisons, which can be estimated by a Kalman filter. Böckenholt, U. and Dillon, W. (1997) suggest a logit model for constant sum paired comparison data, which can be estimated by an EM algorithm. Another approach is given by Cattelan et al (2013) who define a cumulative logit model using an exponentially weighted moving average method.
We instead propose a simple log-linear model which can be fitted (and also be checked) within the framework of generalized linear models (GLMs). For *longitudinal* paired comparison data, we consider the pattern of responses of a single paired comparison response over time. This model allows dependence between time periods to be specified.
Therefore we are interested in discussing a simple log-linear paired comparison model (LLBT model) for repeated observations, following Grand et al (2013) who proposed a log-linear model for repeated observations with temporal dependency parameters with an application to two time points. In longitudinal paired comparison studies individuals are asked to repeatedly make decisions among a set of objects in $\binom{J}{2}$ different comparisons at each time point.

This model assumes independence between paired comparisons within a time point. This assumption is common for many Bradley-Terry models but may be questioned. Dependencies may arise when the responses to a paired comparison are affected by the responses given to other paired comparisons. In the framework of log-linear Bradley-Terry models we could use the paired comparison pattern structure of Dittrich et al (2002), which allows *between-comparison dependencies* to be included and apply this to longitudinal paired

comparison data. However, this would only work for a small number of objects and time points. The discussion section of this paper explains this in more detail. In our approach we therefore assume independence between the paired comparisons. We build the model by considering one comparison (jk) repeated over time. Within this comparison we consider the temporal pattern of responses. We should also expect that a judge who answers in a given paired comparison in a particular way at one time point is more likely to make the same judgement at successive time points. To check this and other temporal dependence assumptions we incorporate for each paired comparison *within-comparison dependency parameters* of a first and second order Markovian structure into the model for repeated observations (LLBTR model). We also want to show how the LLBTR model can be extended by the inclusion of parameters which measure the amount of change of the preference parameters over time. An advantage of this LLBTR approach will be that it is suitable for analysing a relative large number of objects over time. For example, for a large sample size 15 objects resulting in 105 paired comparisons with total first and second order dependency parameters for all paired comparisons over three time points can be modelled.

2 The log-linear Bradley-Terry model for repeated observations

2.1 The log-linear Bradley-Terry model (LLBT)

In each paired comparison (jk) there are two possible responses: preference for object j or preference for object k, which can be interpreted as the realization of the discrete random variable Y_{jk}, where:

$$Y_{jk} = \begin{cases} 1 & \text{if object } j \text{ is preferred over object } k\ , \\ -1 & \text{if object } k \text{ is preferred over object } j\ . \end{cases}$$

The J objects are indexed by j and k and object pairs indexed by (jk), with $j < k\,(j = 1, 2, \ldots, J-1; k = 2, 3, \ldots, J)$.

Let $p_{jk(1)}$ be the probability of preferring object j in the comparison (jk) and $p_{jk(-1)}$ the probability of preferring object k. The Bradley-Terry (BT) model (Bradley and Terry, 1952) defines the probability that object j is preferred over object k, $P(Y_{jk} = 1|\pi_j, \pi_k) = p_{jk(1)} = \frac{\pi_j}{\pi_j + \pi_k}$. The πs are the so called worth parameters of the objects. The worth parameters are non-negative and restricted (with $\sum_j \pi_j = 1$) so that they locate the objects on a latent preference scale between zero and one. The probability for the preference of object k compared to object j is: $P(Y_{jk} = -1|\pi_j, \pi_k) = p_{jk(-1)} = \frac{\pi_k}{\pi_j + \pi_k}$.

In general, following Sinclair (1982), these probabilities can be reformulated as

$$P(Y_{jk} = y_{jk}|\pi_j, \pi_k) = p_{jk(y_{jk})} = c_{jk} \left(\frac{\sqrt{\pi_j}}{\sqrt{\pi_k}} \right)^{y_{jk}}, \qquad (1)$$

with $y_{jk} \in \{-1, 1\}$. c_{jk} is a normalizing constant which is definied by ${c_{jk}}^{-1} = \sqrt{\pi_j/\pi_k} + \sqrt{\pi_k/\pi_j}$ and does not depend on y_{jk}.

Over a sample of judges, the total number of comparisons made between two objects of a given paired comparison (jk) is represented by N_{jk}, which is the sum of the number of preferences of object j, $n_{jk(1)}$, and of the number of preferences of object k, $n_{jk(-1)}$: $N_{jk} = n_{jk(1)} + n_{jk(-1)}$. The expected number of preferences for object j and object k, respectively are given by $m_{jk(1)} = N_{jk}p_{jk(1)}$ and $m_{jk(-1)} = N_{jk}p_{jk(-1)}$.

For each paired comparison (jk) the log-linear representation of the Bradley-Terry model as formulated in (1) is (cf. Dittrich et al, 1998),

$$\ln m_{jk(1)} = \mu_{jk} + \lambda_j - \lambda_k \qquad \text{and} \qquad \ln m_{jk(-1)} = \mu_{jk} - \lambda_j + \lambda_k \ , \qquad (2)$$

where λ_j is the (preference) parameter for object j and $\lambda_j = \frac{1}{2}\ln \pi_j$ or $\exp(2\lambda_j) = \pi_j$. The nuisance parameters μ_{jk} are normalizing constants. The model can be fitted as a Poisson log-linear model with y-variate $n_{jk(1)}$ and a set of nuisance parameters μ_{jk}. For identifiability we set λ_J to be zero.

2.2 The LLBT model for repeated observations (LLBTR)

The log-linear Bradley-Terry model (LLBT) defined in (2) is used for modelling single responses made by judges in a given paired comparison (jk) at one time point. Extending this model, we now consider paired comparisons repeated over time points t, $t = 1, 2, \ldots, T$. There are $L = 2^T$ possible response patterns $\mathcal{T}_{\ell|T}$=$(y_{jk1}, \ldots, y_{jkt}, \ldots, y_{jkT})$, $\ell = 1, 2, \ldots, L$, in a given paired comparisons (jk). In each paired comparison (jk) there are two possible responses at each time point t: preference for object j at time point t, which is denoted by $Y_{jkt} = 1$ or preference for object k at time point t, denoted by $Y_{jkt} = -1$.

Example: For three time points (T=3) we get eight different response patterns, i.e.

$$\begin{aligned}
\mathcal{T}_{1|3} &= (Y_{jk1} = \ \ 1, Y_{jk2} = \ \ 1, Y_{jk3} = \ \ 1) = (\ \ 1 \ \ \ 1 \ \ \ 1)\\
\mathcal{T}_{2|3} &= (Y_{jk1} = -1, Y_{jk2} = \ \ 1, Y_{jk3} = \ \ 1) = (-1 \ \ \ 1 \ \ \ 1)\\
\mathcal{T}_{3|3} &= (Y_{jk1} = \ \ 1, Y_{jk2} = -1, Y_{jk3} = \ \ 1) = (\ \ 1 -1 \ \ \ 1)\\
\mathcal{T}_{4|3} &= (Y_{jk1} = -1, Y_{jk2} = -1, Y_{jk3} = \ \ 1) = (-1 -1 \ \ \ 1)\\
\mathcal{T}_{5|3} &= (Y_{jk1} = \ \ 1, Y_{jk2} = \ \ 1, Y_{jk3} = -1) = (\ \ 1 \ \ \ 1 -1)\\
\mathcal{T}_{6|3} &= (Y_{jk1} = -1, Y_{jk2} = \ \ 1, Y_{jk3} = -1) = (-1 \ \ \ 1 -1)\\
\mathcal{T}_{7|3} &= (Y_{jk1} = \ \ 1, Y_{jk2} = -1, Y_{jk3} = -1) = (\ \ 1 -1 -1)\\
\mathcal{T}_{8|3} &= (Y_{jk1} = -1, Y_{jk2} = -1, Y_{jk3} = -1) = (-1 -1 -1)
\end{aligned}$$

The response pattern $\mathcal{T}_{1|3}$, for example, means that object j is preferred compared to object k at time points $t = 1$, $t = 2$ and $t = 3$ in the comparison (jk), $(Y_{jk1} = 1, Y_{jk2} = 1, Y_{jk3} = 1)$ or in shortened form (111). To generate all possible response patterns for $t = 1, \ldots, T$ time points, we use a pre-defined standard order of temporal within comparison patterns where the last time period varies most slowly.

For each paired comparison (jk) the probability of observing a certain response pattern $\mathcal{T}_{\ell|T}$ for T time points (cf. Grand et al, 2013) is given by

$$p_{jk(y_{jk1},\ldots,y_{jkT})} = c^*_{jk} \prod_{t=1}^{T} \left(\frac{\sqrt{\pi_{jt}}}{\sqrt{\pi_{kt}}} \right)^{y_{jkt}} , \tag{3}$$

where c^*_{jk} is a normalizing constant to make the probabilities sum to one. The parameter π_{jt} is the worth parameter of object j at time t. It is assumed that the decisions between the judges are independent and that each judge makes decisions in all paired comparisons at each time point. At this stage we assume in a given comparison independence of the decisions between two time points. This is the standard independence model.

The number of judges N_{jk} comparing the objects j and k over T time points, is: $N_{jk} = \sum_{\ell=1}^{L} n_{jk(\mathcal{T}_{\ell|T})}$. Let the number of times where response pattern $\mathcal{T}_\ell$ occurs over T time points in the comparison (jk), $n_{jk(\mathcal{T}_{\ell|T})}$, be the random variable. Then the $n_{jk(\mathcal{T}_{\ell|T})}$s are multinomially distributed with N_{jk} and with probabilities $p_{jk(\mathcal{T}_{\ell|T})}$. However, a multinomial model can be fitted as a (conditional) Poisson log-linear model (Aitkin et al, 2009) providing that the sum of the expected cell counts for a particular comparison (jk) is constrained to N_{jk}. This is achieved by adding a set of nuisance parameters μ_{jk} to the model. The expectations $m_{jk(\mathcal{T}_{\ell|T})}$ of $n_{jk(\mathcal{T}_{\ell|T})}$ are given by $m_{jk(\mathcal{T}_{\ell|T})} = N_{jk} p_{jk(\mathcal{T}_{\ell|T})}$.

In general, for T time points, the LLBT model for repeated observations (LLBTR) for each response pattern in each paired comparison (cf. Grand et al, 2013) is defined by:

$$\ln m_{jk(y_{jk1},\ldots,y_{jkT})} = \mu_{jk} + \sum_{t=1}^{T} y_{jkt} \left(\lambda_{jt} - \lambda_{kt} \right) , \tag{4}$$

where $y_{jkt} \in \{-1, 1\}$ is the observed response made in the paired comparison (jk) at time point t and λ_{jt} is the object or preference parameter of object j at time t. The μ_{jk} parameters for the comparisons (jk) are normalizing constants. Note that this LLBTR (independence) model is equivalent to fitting a separate LLBT model for each time point. For identifiability we set λ_{Jt} to be zero for all t. The LLBTR model (4) has 2^T equations – one equation for one of the possible response patterns $\mathcal{T}_{\ell|T}$ – for each of the $\binom{J}{2}$ paired comparisons (jk).

The worth parameters π of equation (3) are obtained, with the requirement that $\sum_j \pi_{jt} = 1$, by

$$\pi_{jt} = \frac{\exp(2\lambda_{jt})}{\sum_{j=1}^{J} \exp(2\lambda_{jt})} . \tag{5}$$

2.2.1 Temporal within-comparison dependencies

LLBTR model with Markovian structure of first and second order

In each paired comparison (Y_{jkt}) at each given time point t, $t = 1, 2, \ldots, T$, there are two possible responses which can be observed: $y_{jkt(1)}$, for the preference of object j and $y_{jkt(-1)}$, for the preference of object k at time point t. The two possible realizations of Y_{jkt} can be thought of as the *states* of a Markov chain.

For each paired comparison (jk) a Markov chain of first order is defined by

$$P\left(Y_{jkt} = y_{jkt} | Y_{jk,t-1} = y_{jk,t-1}, \ldots, Y_{jk0} = y_{jk0}\right) = P\left(Y_{jkt} = y_{jkt} | Y_{jk,t-1} = y_{jk,t-1}\right),$$

which means that only the previous response at time $t-1$ has an influence on the response at time t for a given paired comparison (jk) irrespective of all the responses in the past. A Markov chain of second order is given by

$$P\left(Y_{jkt} = y_{jkt} | Y_{jk,t-1} = y_{jk,t-1}, Y_{jk,t-2} = y_{jk,t-2}, \ldots, Y_{jk0} = y_{jk0}\right) = P\left(Y_{jkt} = y_{jkt} | Y_{jk,t-2} = y_{jk,t-2}, Y_{jk,t-1} = y_{jk,t-1}\right),$$

where now the response at time point t depends on the two previous responses at time $t-2$ and $t-1$ in a given paired comparison (jk). Lindsey (1992), for example, shows how to check the assumptions of a first and second order Markov chain by fitting a log-linear model.

Temporal within-comparison dependencies of first order can be incorporated into the LLBTR model by the interaction parameters $\zeta_{jk|t-1,t}$:

$$\ln m_{jk(y_{jk1},\ldots,y_{jkT})} = \mu_{jk} + \sum_{t=1}^{T} y_{jkt}(\lambda_{jt} - \lambda_{kt}) + \sum_{t=2}^{T} y_{jk,t-1} y_{jkt} \zeta_{jk|t-1,t} \ . \quad (6)$$

The $\zeta_{jk|t-1,t}$ parameters represent possible interactions between the responses at two consecutive time points $t-1$ and t of a certain comparison (jk). For each of the $\binom{J}{2}$ paired comparisons we can specify the temporal dependence structure by a set of $T-1$ ζs. When $\zeta_{jk|t-1,t} = 0$ for each paired comparison, we get the independence model (4).

Extending model (6) by $\zeta_{jk|t-2,t}$, we get the LLBTR model with second order within-comparison dependencies

$$\begin{aligned} \ln m_{jk(y_{jk1},\ldots,y_{jkT})} =& \mu_{jk} + \sum_{t=1}^{T} y_{jkt}(\lambda_{jt} - \lambda_{kt}) + \sum_{t=2}^{T} y_{jk,t-1} y_{jkt} \zeta_{jk|t-1,t} + \\ & \sum_{t=3}^{T} y_{jk,t-2} y_{jkt} \zeta_{jk|t-2,t} \ , \end{aligned} \quad (7)$$

where now the $\zeta_{jk|t-2,t}$s represent the interaction between the decisions $Y_{jk,t-2}$ and Y_{jkt} at time point $t-2$ and t. The λ_{Jt}'s are again set to be zero for all t.

2.2.2 Modelling change

To model possible change effects for the J object parameters or preference parameters over time, we could reparameterise the LLBTR model and incorporate temporal change parameters δ_j for each of the J object parameters. These parameters represent the amount of change of the object parameters (the λs) from time point $t = 1$ to another (see also Glickman, 1993).

Example: Assuming we have only two time points, the δ parameters for the change of the preference parameters from time point $t = 1$ to time point $t = 2$ are defined by: $\delta_{j|2,1} = \lambda_{j2} - \lambda_{j1}$. By replacing the preference parameters λ_{j2} of model (4) at time point $t = 2$ with $\lambda_{j1} + \delta_{j|2,1}$, we get the LLBTR model for two time points with change parameters δ:

$$\ln m_{jk(y_{jk1},y_{jk2})} = \mu_{jk} + y_{jk1}(\lambda_{j1} - \lambda_{k1}) + y_{jk2}(\lambda_{j1} + \delta_{j|2,1} - \lambda_{k1} - \delta_{k|2,1}) \ . \quad (8)$$

More generally, the LLBTR model with temporal change parameters $\delta_{j|t,1}$ which indicate the amount of change of the object parameters λ between the first time point (the reference time point, $t = 1$) and time point t is defined by

$$\ln m_{jk(y_{jk1},\ldots,y_{jkT})} = \mu_{jk} + y_{jk1}(\lambda_{j1} - \lambda_{k1}) + \sum_{t=2}^{T} y_{jkt}(\lambda_{j1} + \delta_{j|t,1} - \lambda_{k1} - \delta_{k|t,1}) \ . \quad (9)$$

For two time points we get one change parameter and for $T > 2$ we get $T - 1$ parameters $\delta_{j|t,1}$ for each of the J objects. For identifiability, $\delta_{J|t,1}$ is set to zero for all t. These parameters can be interpreted as conditional log-odds. In the simple case of only two time points $T = 2$ and evenly matched objects j and k at time point 1 (i.e. $\lambda_{j1} = \lambda_{k1}$), the log-odds for the preference of object j compared to object k at time point 2 is

$$\begin{aligned} \ln \frac{p_{jk(y_{jk1},y_{jk2(\ \ 1)})}}{p_{jk(y_{jk1},y_{jk2(-1)})}} &= \ln m_{jk(y_{jk1},y_{jk2(1)})} - \ln m_{jk(y_{jk1},y_{jk2(-1)})} \\ &= 2(\delta_{j|2,1} - \delta_{k|2,1}) \ . \end{aligned}$$

2.3 Interpretation of the parameters of model (6) and (7)

The object parameters λ_j and the temporal within-comparison dependence parameters $\zeta_{jk|t-1,t}$ can be interpreted as conditional log-odds. For example, the log-odds in favour of object j compared to object k at time point T can easily be calculated given all previous decisions at time points $t = 1, \ldots, T-1$. Fitting a model (7) with a second order Markovian structure the log-odds in favour of object j in the comparison with object k at time point T depends on the values given for $y_{jk,T-1}$ and $y_{jk,T-2}$. There are four possible conditional log-odds:

$$\ln \frac{p_{jk(y_{jkT(\ \ 1)}|y_{jk,T-1(\ \ 1)},y_{jk,T-2(\ \ 1)},\cdots)}}{p_{jk(y_{jkT(-1)}|y_{jk,T-1(\ \ 1)},y_{jk,T-2(\ \ 1)},\cdots)}} = 2(\lambda_{jT}-\lambda_{kT})+2(\ \ \zeta_{jk|T-1,T}+\zeta_{jk|T-2,T})$$

$$\ln \frac{p_{jk(y_{jkT(\ \ 1)}|y_{jk,T-1(-1)},y_{jk,T-2(\ \ 1)},\cdots)}}{p_{jk(y_{jkT(-1)}|y_{jk,T-1(-1)},y_{jk,T-2(\ \ 1)},\cdots)}} = 2(\lambda_{jT}-\lambda_{kT})+2(-\zeta_{jk|T-1,T}+\zeta_{jk|T-2,T})$$

$$\ln \frac{p_{jk(y_{jkT(\ \ 1)}|y_{jk,T-1(\ \ 1)},y_{jk,T-2(-1)},\cdots)}}{p_{jk(y_{jkT(-1)}|y_{jk,T-1(\ \ 1)},y_{jk,T-2(-1)},\cdots)}} = 2(\lambda_{jT}-\lambda_{kT})+2(\ \ \zeta_{jk|T-1,T}-\zeta_{jk|T-2,T})$$

$$\ln \frac{p_{jk(y_{jkT(\ \ 1)}|y_{jk,T-1(-1)},y_{jk,T-2(-1)},\cdots)}}{p_{jk(y_{jkT(-1)}|y_{jk,T-1(-1)},y_{jk,T-2(-1)},\cdots)}} = 2(\lambda_{jT}-\lambda_{kT})+2(-\zeta_{jk|T-1,T}-\zeta_{jk|T-2,T})\ .$$

We can see that the log-odds in favour of object j in the comparison (jk) at time point T depends on the object parameters λ_{jT} and λ_{kT} and there is also a possible effect from the decision made at $T-1$ and $T-2$ represented by the parameters $\zeta_{jk|T-1,T}$ and $\zeta_{jk|T-2,T}$.

The within-comparison dependence parameters can also be interpreted as conditional log-odds ratios. For example, the log-odds ratio of making a consistent decision (i.e. $(\ \ 1\ \ 1)$, $(-1\ -1)$) at time points $T-1$ and T (or $T-2$ and T in the case of second order dependencies) against an inconsistent decision (i.e. $(\ \ 1\ -1)$, $(-1\ \ 1)$), can be calculated by conditioning with respect to all decisions at the previous time points. The conditional log-odds ratio for making consistent decisions compared to inconsistent decisions at time point $T-1$ and T is:

$$\ln \frac{p_{jk(y_{jk,T-1(\ \ 1)},y_{jkT(\ \ 1)}|y_{jk,T-2},\cdots)} \cdot p_{jk(y_{jk,T-1(-1)},y_{jkT(-1)}|y_{jk,T-2},\cdots)}}{p_{jk(y_{jk,T-1(\ \ 1)},y_{jkT(-1)}|y_{jk,T-2},\cdots)} \cdot p_{jk(y_{jk,T-1(-1)},y_{jkT(\ \ 1)}|y_{jk,T-2},\cdots)}} =$$
$$\ln m_{jk(111)} + \ln m_{jk(1-1-1)} - \ln m_{jk(11-1)} - \ln m_{jk(1-11)} = 4\zeta_{jk|T-1,T}\ .$$

2.4 Attrition in longitudinal paired comparisons

The proposed LLBTR model is appropriate for analysing longitudinal paired comparison data where individuals respond at all time points. In practice, however, individuals often choose not to participate at all time points, dropping out of one or more sweeps of the study and for these individuals missing values across all paired comparisons at one or more time points will be recorded. If there are individuals with incomplete response patterns over T time points (i.e. with missing values) the simplest way is to remove such incomplete patterns; this is known as a complete case analysis, However, where there a large number of individuals fail to respond at one or more time points, a complete case analysis loses information. Additionally, the complete case method is problematic if subjects with incomplete response patterns systematically differ from subjects with complete patterns.

There are two approaches to the attrition problem which we will consider here. The first is to use all of the recorded data, but assume that the missing structure is uninformative on the estimates. This is known as a full information

maximum likelihood (FIML) analysis. We proceed by building on our earlier model, and identify all possible missing patterns present in the data in terms of whether the individual responded or not at the time points of the survey. For example, for $T = 3$, there are eight possible missingness patterns `(x x x)`, `(x NA x)`, `(x x NA)`, `(x NA NA)`, `(NA x x)`, `(NA NA x)`, `(NA x NA)` and `(NA NA NA)`, where `NA` represents a missing response and `X` an observed response. In practice, there may be fewer missingness patterns, as the number of individuals who fail to respond at the first time point may be zero. Let these P observed missingness patterns be indexed by p. We note that the number of response patterns for each missingness pattern will vary. For example, for $T = 3$, for the complete data pattern `(x x x)` there are eight response patterns as previously identified, but for the pattern `(x NA NA)` there are only two possible responses `(1 NA NA)` and `(-1 NA NA)`. We now operationalise this model by extending the definition of Y in section 7 and 8. We now additionally define $Y_{jkt} = 0$ where a response at a specific time point t is missing. Then missing responses do not contribute to either the estimate of the object parameter at that time point, or to any interaction involving that missing response. In this way, the full set of observed data can be used.

The model now becomes:

$$\begin{aligned}\ln m_{jkp(y_{jk1},\ldots,y_{jkT})} =& \mu_{jkp} + \sum_{t=1}^{T} y_{jkt}(\lambda_{jt} - \lambda_{kt}) + \sum_{t=2}^{T} y_{jk,t-1} y_{jkt} \zeta_{jk|t-1,t} + \\ & \sum_{t=3}^{T} y_{jk,t-2} y_{jkt} \zeta_{jk|t-2,t} \ ,\end{aligned} \tag{10}$$

where the μ_{jkp} is now additionally indexed by p to allow the marginal totals within each paired comparison for each missingness pattern to be reproduced. Note that the FIML model without dependencies is equivalent to fitting separate LLBT models to each time point. However, the FIML LLBTR model *with dependencies* gives fitted values which are closer to the observed counts.

The assumption of uninformative missing may however be questionable if subjects with incomplete response patterns systematically differ from subjects with complete patterns. A simple approach for treating this missing issue is to use the missingness pattern as an additional categorical covariate with P levels ($p = 1, 2 \ldots, P$), indicating the missingness behaviour of individuals over time. This can then be interacted with the object effects in the model, allowing the estimation of different object parameters for each missing data pattern. A more complex model is to also allow the dependence terms to depend on the missingness pattern. If there are a large number of missingness patterns, they can be simplified by summarizing the missingness patterns on the basis of a particular criterion, e.g. : a two level factor contrasting individuals with complete vs. incomplete missingness patterns, or a continuous covariate counting the number of time points responded to. This is known as a pattern-mixture model (see Hedeker and Gibbons, 1997).

3 Course design preferences in statistics

Students attending a statistics course in 2011 and 2012 given by the same instructor were repeatedly surveyed in a paired comparison experiment of preferences relating to possible course designs (objects). The students were asked the same paired comparisons at four consecutive time points or sweeps (at the beginning of the course, before the first written test, after the first written test in statistics and at the end of the course) in class but they also had the opportunity to fill out the questionnaire online. The six paired comparisons were selections of two of the four possible course designs - the objects in this study. The course designs were: "*complete direct instruction*" (course design 1), "*Partial direct instruction (self-initiated content preparation combined with direct instruction)*" (course design 2), "*partial self-regulated learning (e-learning with discussion of the solution)*" (course design 3) and "*complete self-regulated learning (e-learning alone)*" (course design 4). Each of the course designs consisted of core didactic elements (content, examples, solutions, discussion) where some or at least one of them can be fulfilled in class by the instructor and/or individually and independently by using the e-learning platform learn@WU (for details see Grand et al, 2013).

Our initial requirement for our analysis is that students had to have responded at the first time point. Only six students failed to take part at this stage and these students were excluded. In addition, the number of students who responded at time point 4 was very small. We therefore focused our analysis on the course design preferences of students over three consecutive time points (at the beginning of the course ($t = 1$), before the first test ($t = 2$) and after the first test ($t = 3$) in statistics). The analysis was carried out in R (R Development Core Team, 2013) using the packages prefmod (Hatzinger, 2012) to set up the design matrices for analysis, and gnm (Turner and Firth, 2012) to fit the models. The participation of respondents over the three time points of the sample is shown in Table 1, and the preferences of students to individual comparisons for the three time points is given in Table 2. Table 1 shows that there is substantial attrition after the first time point which will need to be taken account of in the analysis. In addition, Table 2 shows that course design 1 appears to be the most popular choice, but it is hard to determine the ordering of the other designs.

Table 1 Participation over the three time points

	time point 1	time point 2	time point 3	no. of respondents
	x	x	x	171
	x	x	NA	126
	x	NA	x	22
	x	NA	NA	347
no. of respondents	666	297	193	666

Note: Participation is denoted by x and non-participation by NA.

Table 2 Responses for each comparison at each time point

	time point 1 total (n=666)		**time point 2** total (n=297)		**time point 3** total (n=193)	
Paired comparison	1st preferred	2nd preferred	1st preferred	2nd preferred	1st preferred	2nd preferred
(12)	384	282	162	135	98	95
(13)	493	173	200	97	122	71
(23)	506	160	229	68	147	46
(14)	532	134	229	68	147	46
(24)	579	87	259	38	166	27
(34)	553	113	243	54	165	28

3.1 FIML analysis with Markov dependencies

Our analysis starts with a full information maximum likelihood (FIML) approach where the full data (i.e. all observed data) are modelled. For each paired comparison (jk) for time points 2 and 3 we can observe three possible outcomes: $y_{jk} = 1, y_{jk} = -1$ and $y_{jk} = 0$, whereas for time point 1 there are no missing responses and there are only two possible outcomes $y_{jk} = 1$ and $y_{jk} = -1$. Thus, there are $2 \times 3 \times 3 = 18$ possible response patterns for each paired comparison (jk). We then fitted an initial FIML LLBTR model (model 1 in Table 3).

The next stage was to try to simplify the 18 dependence parameters. One of the second order dependence parameters ($\zeta_{14|1,3}$) was not significant and was set to zero. The other second order dependence parameters were similar and were set equal to each other ($\zeta_C = \zeta_{12|1,3} = \zeta_{13|1,3} = \zeta_{23|1,3} = \zeta_{24|1,3} = \zeta_{34|1,3}$). In examining the twelve first order dependencies, we found that they could be simplified to two values, the first value equating two of the $t1, t2$ dependencies and all of the $t2, t3$ dependencies ($\zeta_A = \zeta_{24|1,2} = \zeta_{34|1,2} = \zeta_{12|2,3} = \zeta_{13|2,3} = \zeta_{23|2,3} = \zeta_{14|2,3} = \zeta_{24|2,3} = \zeta_{34|2,3}$) and the second equating the remaining four $t1, t2$ dependencies ($\zeta_B = \zeta_{12|1,2} = \zeta_{13|1,2} = \zeta_{23|1,2} = \zeta_{14|1,2}$). This became our model 2, with simplified dependencies. Table 3 shows that the deviance increase in moving from model 1 to model 2 is small and not significant (Δdeviance = 6.97 on 15 df, p=0.95).

The first column of Table 4 shows the parameter estimates for the FIML model with simplified dependence parameters (model 2). We can observe that the spread of the λ parameters is large for the first time point, but narrows for time points 2 and 3. Figure 1 shows the derived worth values for model 2 for the three time points together with the 95% confidence intervals. These confidence intervals were calculated using the delta method (Agresti, 2013, p72-75) using the R package msm (Jackson, 2011). The worth estimate for course design 1 (direct instruction) can be seen to steadily decline whereas the worth estimate for course design 2 (self-initiated content preparation) is stable, and the worths for course design 3 (e-learning with discussion) increases steadily. Over the three time points, the rank order of course design 1 and 2

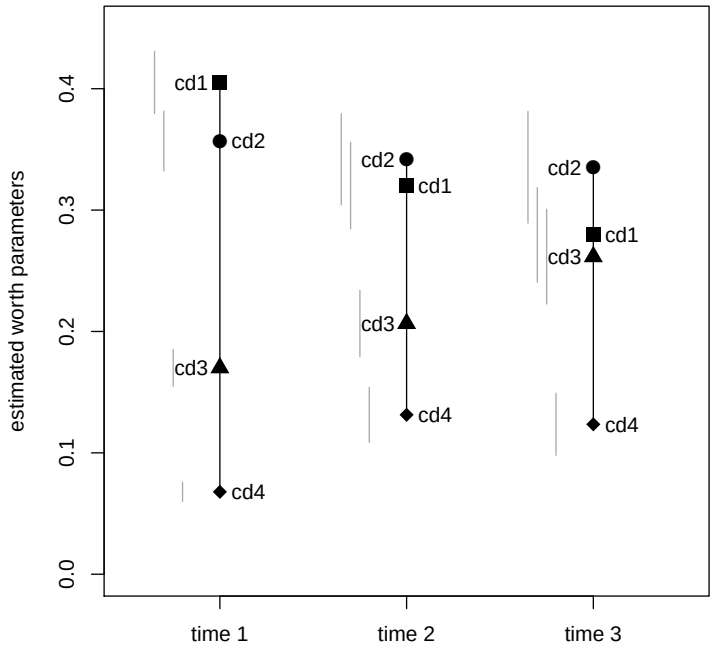

Fig. 1 Worth parameters and 95% confidence intervals for the four course designs (cd) produced by the full information LLBTR model with simplified dependencies (model 2).

changes, with course design 2 becoming the most popular course design in time points 2 and 3, and the gap between cd1 and cd2 further widening by time point 3.

3.2 Allowing for misssingness through the two-group pattern mixture model

So far, we have not distinguished between those students giving complete responses and those giving incomplete responses. However, we were interested in whether the preference parameters of the course designs may be different across these two groups. We therefore compared students showing a complete response pattern with those showing an incomplete pattern, and fitted a LLBTR model over three time points, additionally including an interaction between the complete/incomplete participation factor and the object parameters at each time point. We refer to this model as the two-group pattern mixture model.

In moving from the FIML model (model 2) to the two group pattern mixture model (model 3), there is a significance deviance change of 24.80 on 9

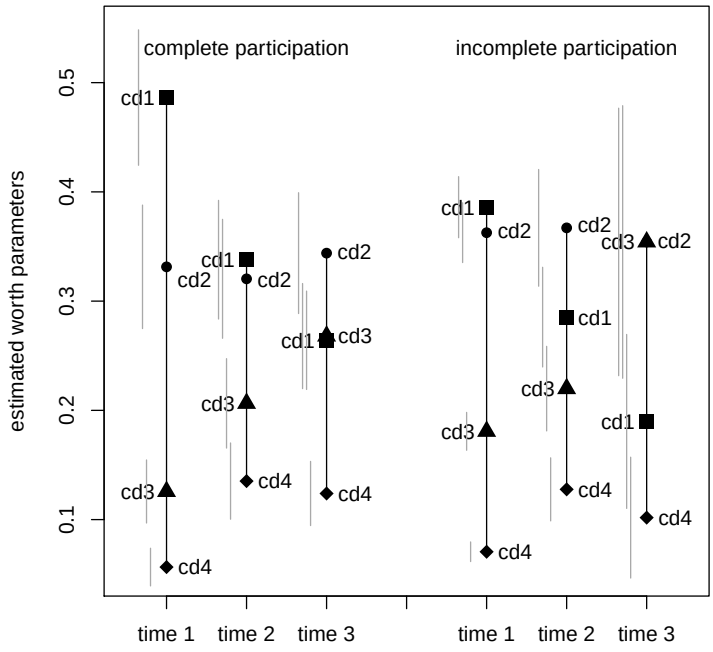

Fig. 2 Worth parameters and 95% confidence intervals for the two-group pattern-mixture model showing worths for students with complete and incomplete participation over the three time points (model 3).

df (p=0.003) (see Table 3). This shows that there are large and significant differences between the worths for the complete responders and those for the incomplete responders. The second column of Table 4 shows the parameter estimates for this model. In examining the object parameters, we notice that the spread in the time point 1 parameters for the complete participation group widens compared with model 2. The object parameters for the incomplete participation group are interpreted as additional effects on the object parameters for the complete group. We note that the object parameter for cd1 is large and negative, indicating that course design 1 is less preferred in the incomplete group.

To examine the *changes* in the object parameters over time, we reparameterise model 3 and fit the change model (model 3C). This model has the same deviance and degrees of freedom as model 3, but the λs at time points 2 and 3 are replaced by δs, following the model described in Section 2.2.2. The third column of Table 4 gives the estimates from this model. This allows us to observe that the change parameters for the complete participants are significant, whereas the change parameters for the incomplete participants are not.

Table 3 Deviances of hierarchical models with change in deviances

LLBTR models	**deviance**	**df**	**deviance change** from previous model	**df change**	**p-value**
1. Full information ML model (1 group) all pairwise temporal dependencies	178.20	57			
2. Full information ML model (1 group) simplified temporal dependencies	185.17	72	6.97	15	0.958
3. Pattern-mixture model (2 groups) complete vs incomplete paticipation simplified temporal dependencies	160.37	63	24.80	9	0.003

In other words, the complete participants are changing substantially over time in the course design preferences, and the incomplete participants are changing in general in the same way, with nearly no *additional* change over time.

Figure 2 shows the model 3 worth estimates and 95% confidence intervals separately for the complete and incomplete participants. For the complete participants, the worths of the objects change between the start of the course (t=1) and the time after having taken the first test (t=3). At the first time point, the most preferred course design is that of direct instruction (cd1) followed by cd2 - the self-initiated content preparation with direct instruction. These two course designs are preferred over the course designs e-learning with discussion (cd3) and e-learning only (cd4). The worths for the course designs become more similar over time, with cd2 overtaking cd1 to become the most preferred design by time point 3. Over all time points the e-learning only course design is the least preferred course design. For the incomplete participants, we can see that the designs cd1 and cd2 are very similar at the first time point. The design cd2 becomes the most preferred design at time point 2, and cd2 and cd3 (e-learning with discussion) are equally preferred at time point 3. Again, e-learning without discussion (cd4) is the least preferred design over all three time points. For both sets of respondents, the worth parameters of the direct instruction design (cd1) falls over the three time points.

We now examine the dependency parameters for the pattern mixture model. We note that there is a positive second-order dependency for nearly all paired comparisons between time points 1 and 3 and over all possible consecutive time points (t1-t2) and (t2-t3). One second order dependency however is set to zero - for paired comparison (14). Between time point 1 and 2, the within-comparison dependence parameter for the comparisons (24) and (34) ($\zeta_A = 0.528$) is larger than the parameter for all other paired comparisons, ($\zeta_B = 0.312$). For time points 2 and 3, the dependencies are large for all paired comparisons and are set to ζ_A. We can interpret these parameters in terms of conditional odds. For example, the conditional odds ratio for making consistent decisions at time

Table 4 Estimates of LLBTR models: full information model with simplified temporal dependencies (model 2), pattern-mixture model with simplified temporal dependencies (model 3) and change model (model 3C).

	full information model (Model 2) dev.: 185.17, df: 72		**pattern-mixture model (Model 3)** dev.: 160.37, df: 63		**change model (Model 3C)** dev.: 160.37, df: 63		
	estimate	s.e.	estimate	s.e.		estimate	s.e.
λ_{11}	0.893	0.037	1.076	0.091	λ_{11}	1.076	0.091
λ_{21}	0.830	0.037	0.884	0.095	λ_{21}	0.884	0.095
λ_{31}	0.459	0.035	0.400	0.087	λ_{31}	0.400	0.087
λ_{41}	0.000	NA	0.000	NA	λ_{41}	0.000	NA
λ_{12}	0.446	0.058	0.458	0.086	$\delta_{1\|2,1}$	-0.618*	0.139
λ_{22}	0.478	0.060	0.431	0.091	$\delta_{2\|2,1}$	-0.453*	0.148
λ_{32}	0.226	0.056	0.211	0.085	$\delta_{3\|2,1}$	-0.188	0.139
λ_{42}	0.000	NA	0.000	NA	$\delta_{4\|2,1}$	0.000	NA
λ_{13}	0.408	0.068	0.379	0.078	$\delta_{1\|3,1}$	-0.697*	0.126
λ_{23}	0.499	0.073	0.510	0.084	$\delta_{2\|3,1}$	-0.374*	0.136
λ_{33}	0.375	0.067	0.386	0.079	$\delta_{3\|3,1}$	-0.014	0.128
λ_{43}	0.000	NA	0.000	NA	$\delta_{4\|3,1}$	0.000	NA
ζ_A	0.526	0.036	0.528	0.036	ζ_A	0.528*	0.036
ζ_B	0.317	0.036	0.312	0.037	ζ_B	0.312*	0.037
ζ_C	0.334	0.044	0.338	0.045	ζ_C	0.338*	0.045
$\zeta_{14\|1,3}$	0.000	NA	0.000	NA	$\zeta_{14\|1,3}$	0.000	NA
λ_{11}:inc	-	-	-0.226	0.100	λ_{11}:inc	-0.226*	0.100
λ_{21}:inc	-	-	-0.066	0.103	λ_{21}:inc	-0.066	0.103
λ_{31}:inc	-	-	0.071	0.095	λ_{31}:inc	0.071	0.095
λ_{41}:inc	-	-	0.000	NA	λ_{41}:inc	0.000	NA
λ_{12}:inc	-	-	-0.056	0.111	$\delta_{1\|2,1}$:inc	0.171	0.164
λ_{22}:inc	-	-	0.097	0.117	$\delta_{2\|2,1}$:inc	0.162	0.173
λ_{32}:inc	-	-	0.060	0.111	$\delta_{3\|2,1}$:inc	-0.010	0.164
λ_{42}:inc	-	-	0.000	NA	$\delta_{4\|2,1}$:inc	0.000	NA
λ_{13}:inc	-	-	-0.067	0.185	$\delta_{1\|3,1}$:inc	0.159	0.215
λ_{23}:inc	-	-	0.113	0.202	$\delta_{2\|3,1}$:inc	0.179	0.234
λ_{33}:inc	-	-	0.238	0.197	$\delta_{3\|3,1}$:inc	0.167	0.226
λ_{43}:inc	-	-	0.000	NA	$\delta_{4\|3,1}$:inc	0.000	NA

The course designs are labelled as follows: 1: complete direct instruction, 2: partial direct instruction , 3: partial self-regulated learning, 4: complete self-regulated learning. The λ and δ parameters correspond to the group of students with a complete response pattern (i.e. the reference group) and the λ:inc and δ:inc terms refer to the group of students showing an incomplete response pattern and are interpreted relative to the complete group. Significance of parameters is shown only for model 3C (* indicates $p < 0.05$ based on the Wald test).

point 2 and 3 (i.e. preferring the first or the second course design at both time points) compared to inconsistent decisions is $exp(4 \times 0.528) = 8.265$, which indicates a strong tendency for stable judgements in all paired comparisons. The conditional odds ratio of consistent decisions for most paired comparisons between time points 1 and 2 is lower ($exp(4 \times 0.312) = 3.483$) indicating slightly more volatility.

3.3 Lack of fit and overdispersion

One final issue is the apparent lack of fit of the model. Model 3 has a deviance of 160.37 on 63 df, and a standard chi-squared goodness of fit test suggests that this model does not fit ($p < 0.001$). However, there are zero counts in the data, with four of the 108 values of y equal to zero, and the use of the goodness of fit test on such data may be suspect. If we believe the goodness of fit test, then one reason for the lack of fit may be that some cells are badly fitted. Examination of the residuals shows two cells with deviance residuals over 3.5 in absolute value. Both residuals are for incomplete participants. The first large residual comes from the response [(14)4 NA NA] and is positive (3.70) and thus underpredicts the number of design 4 preferences in the (14) comparison. The second is negative (-3.72) and comes from the response [(34)4 NA NA], and overpredicts the design 4 preferences in the (34) comparison. This represents evidence of what Dittrich et al (2004) call response bias and can be dealt with by including extra parameters in the model. However, these two observations do not account for all of the overdispersion in the model. Another reason for the lack of fit already discussed may be that there are subsets of respondents with different latent orderings of the four designs. Thus, response may depend on age, gender or exam performance of the student at time 1. The lack of inclusion of such covariates in this analysis will generate overdispersion. One simple method of dealing with overdispersion is to fit a quasi-Poisson model (Ver Hoef and Boveng, 2007) rather than a Poisson model to the cell counts - this is identical to Agresti's method of scaling the standard errors by $\sqrt{X^2/df}$ when fitting a multinomial (Agresti, 2013, p313). In this framework, the parameter estimates remain the same, but an estimated scale parameter allows for the overdispersion. Comparing the quasi-Poisson versions of model 2 to model 3 now no longer gives a significant change of deviance ($F = 1.09$ on 9, 63 df; $p = 0.38$) and thus there is no longer any evidence of preference orderings varying by whether participants are complete or incomplete. We therefore revert to an overdispersed version of model 2. The model with overdispersion will have the same parameter estimates as for model 2 without overdispersion, but the standard errors will be larger. Table 5 shows the revised parameter estimates and standard errors, together with the derived worths and standard errors from model 2 (calculated using the delta method) after allowing for overdispersion. Comparing these standard errors with Table 4 shows an increase in the size of the standard errors of the estimates of around 55%.

Figure 3 shows the final model which takes account of overdispersion. The Figure is identical to Figure 1 except for the 95% confidence intervals, which are wider. Overlapping confidence intervals are an unreliable guide to which objects differ significantly from other objects, and a more formal approach is necessary. It is also useful to compare which objects are significantly different from which other objects for model 2 under the two assumptions of non overdispersion and overdispersion. Table 6 contains two multiple comparison tables under the two assumptions of no overdispersion (fitted using the Poisson

Table 5 Object and worth parameters: Estimates and standard errors from the overdispersed full information LLBTR model (model 2)

model 2 accounting for overdispersion					
	estimate	**s.e.**		**estimate**	**s.e.**
λ_{11}	0.893	0.058	π_{11}	0.405	0.020
λ_{21}	0.830	0.058	π_{21}	0.357	0.020
λ_{31}	0.459	0.055	π_{31}	0.170	0.012
λ_{41}	0.000	NA	π_{41}	0.068	0.006
λ_{12}	0.446	0.090	π_{12}	0.320	0.028
λ_{22}	0.478	0.094	π_{22}	0.342	0.030
λ_{32}	0.226	0.087	π_{32}	0.207	0.022
λ_{42}	0.000	NA	π_{42}	0.131	0.018
λ_{13}	0.408	0.106	π_{13}	0.280	0.031
λ_{23}	0.499	0.114	π_{23}	0.335	0.037
λ_{33}	0.375	0.105	π_{33}	0.262	0.031
λ_{43}	0.000	NA	π_{43}	0.124	0.020

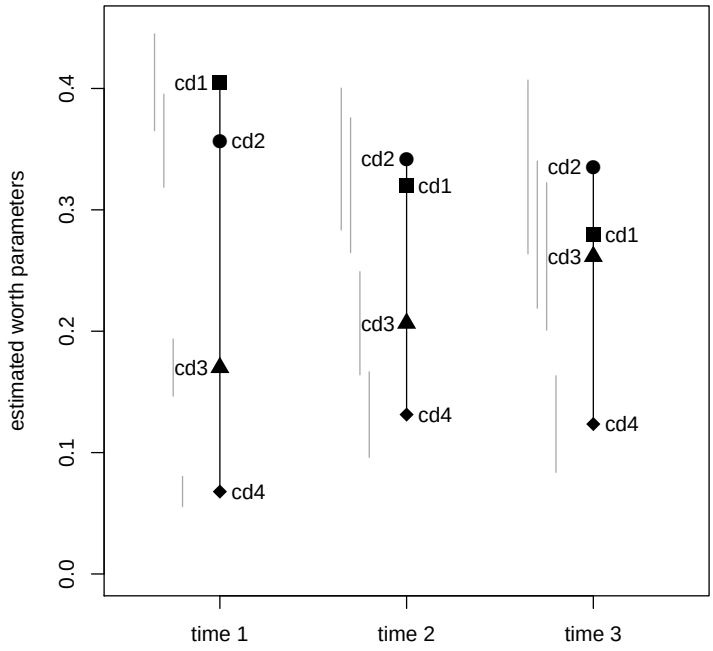

Fig. 3 Worth parameters and 95% confidence intervals for model 2 under the assumption of overdispersion.

Table 6 Multiple comparison tables for objects at each time point under the assumptions of no overdispersion, and overdispersion

	Model 2 no overdispersion					**Model 2 with overdispersion**			
time point 1									
	cd1	cd2	cd3	cd4		cd1	cd2	cd3	cd4
cd1	-				cd1	-			
cd2	*	-			cd2	NS	-		
cd3	*	*	-		cd3	*	*	-	
cd4	*	*	*	-	cd4	*	*	*	-
time point 2									
	cd1	cd2	cd3	cd4		cd1	cd2	cd3	cd4
cd1	-				cd1	-			
cd2	NS	-			cd2	NS	-		
cd3	*	*	-		cd3	*	*	-	
cd4	*	*	*	-	cd4	*	*	*	-
time point 3									
	cd1	cd2	cd3	cd4		cd1	cd2	cd3	cd4
cd1	-				cd1	-			
cd2	NS	-			cd2	NS	-		
cd3	NS	*	-		cd3	NS	NS	-	
cd4	*	*	*	-	cd4	*	*	*	-

A star * indicates a significant difference and NS a non-significant difference at the 5% level..

GLM) and overdispersion (fitted using a quasi-Poisson GLM). The tables were constructed by examining the significance of estimates of the object parameters compared to the reference object using Z-tests. Resetting of the reference object and refitting the model gives identical model fits but alternative parameterisations which allow all comparison pairs to be tested.

There is surprisingly little difference in the two multiple comparison tables. Course designs 1 and 2 are consistently not statistically different to each other under the overdispersion assumption at all three time points but under the no overdispersion model, this consistency disappears, with course designs 1 and 2 differing from each other at time point 1. The only other difference is that a significant difference between course designs 2 and 3 at time point 3 disappears under the overdispersion assumption.

4 Discussion

This paper has presented two methods to handle longitudinal paired comparisons in the presence of attrition - either using a full information maximum likelihood approach or fitting a pattern-mixture model and incorporating the type of missingness pattern as a factor in the model. Both models have incorporated temporal dependency parameters that can be interpreted through log-odds ratios. The choice of which model to fit can be determined through examination of the deviance, although table sparsity may be a problem.

The model can be extended in various ways that are beyond the scope of this paper. We first focus on the dependency structure in the model.

Firstly, it is straightforward to incorporate higher order Markovian dependencies into the LLBTR model if there are more than three time points. Note that the greater the number of time points that are included, the larger the number of possible response patterns. This will lead to an increase in memory usage and a possibly sparse contingency table.

Secondly, an extension to the model in this paper is to combine it with the approach of Dittrich et al (2002), which allows dependencies between paired comparisons. The combination will lead to a concatenation of the paired comparison responses at T time points, producing a single long pattern. For example, in this paper, the pattern vector for three time points is of length 3 with $2^3 = 8$ possible patterns where no attrition is present and $2 \times 3^2 = 18$ where attrition is present at the second and third time point. With four objects and three time points, and using the long pattern form of the combined model, the pattern vector length will be 18, leading to $2^{18} = 262,144$ possible patterns when no attrition is present and $2^6 \times 3^6 \times 3^6 = 3,401,222$ possible patterns where attrition is present. Such combined models which can allow for both kinds of dependency need to be explored further, although it is clear that fitting them will have a high computational load and can therefore only be estimated for a relatively small number of objects.

Finally, we could adopt a marginal approach and fit generalised estimating equation (GEE) models, perhaps using an unstructured form of the correlation matrix to account for time dependencies. A GEE approach naturally takes account of overdispersion in the data through the parametrisation of the correlation matrix. However, model selection is less straightforward particularly with data attrition (Shen and Chen, 2012) and assessment of goodness of fit is problematic.

Other possible extensions to our model involve incorporating temporal dependencies in paired comparison experiments which allow for ties. One route forward is to use a multi-state Markovian structure by incorporating a third state, i.e. the undecided response (see e.g. Lindsey, 1992; Lindsey, 2004). Instead of 8 possible response patterns over three time points in each paired comparison (jk), we would then have $3^3 = 27$ possible response patterns in each paired comparison (jk) and the number of temporal dependence parameters would also increase. Lindsey (1992, 2004), for example, suggested a log linear model for a three-state Markov chain with temporal dependence parameters and their corresponding factor variables. This approach could be taken to build a design structure for a LLBTR model with three response categories and temporal within comparison dependencies, representing possible interactions between the three response categories at consecutive time points of a certain comparison (jk).

Additionally, it is possible to incorporate both object and subject covariates into the model. Dittrich et al (1998) gives details for the cross-sectional model. For longitudinal data, characteristics of the objects may be modelled and may change over time. For example, the number of hours of e-learning,

which is non-zero for course designs 3 and 4, would be an appropriate covariate. Subject covariates can also be included and can also be time-varying. Indeed, incorporation of subject covariates will improve model fit.

We return now to the practical example, and discuss the results of the various models. The model choice depends on whether we believe the goodness of fit test. If we think that the lack of fit is caused by zero cells making the goodness of fit test unreliable, then we would choose model 3, and we identify differences between participants with complete and incomplete responses (Figure 2). If we instead think that the lack of fit is meaningful, and needs to be accounted for, then we would instead choose the overdispersed model 2 (Figure 3). Taking the first route, is it reasonable that incomplete participants are different from complete participants? There is a potential explanation. While students need to attend the examination, there is no strict requirement for them to attend lectures where the questionnaire was delivered. Although students were also followed up by e-mail, the non-responders are students who ignored the questionnaire, either being in the lecture and not completing the task, or not attending at all. Those students who fail to respond to the second or third sweep of the survey could therefore be considered to be less motivated towards the course. These less motivated students are seen to have have less firm views on teaching methods at time point 1 than the more motivated students who filled in all three questionnaires. The analysis therefore provides important results for university teaching policy, suggesting that more e-learning combined with discussion may be beneficial for the less-motivated students as it is preferred after the test has been taken.

The second route, where we account for overdispersion, leads to a simpler model where there are no differences between complete and incomplete participants. In this model (Figure 3), direct instruction is the most preferred course design at the beginning of the statistics course. This may be due to the fact that students are used to direct instruction in the classroom from their school experience. They may also be unaware of other kinds of course design in statistics. However, unlike model 3, there is no significant difference between cd1 and cd2 - the two course designs with a component of direct instruction. The downward change in the worth of course design 1 (direct instruction) from time point 1 to time point 3 (see Table 5) may be due to a lessening of interest in the complete direct instruction design. This, in turn, might be due to an increased demand for flexibility and independence from formal learning (which requires students to learn at a fixed place at a fixed time) while still having the opportunity of discussing solutions and receiving feedback.

Are the results practically as well as statistically significant? Taking the overdispersed model, at all three time points the odds in favour of the course designs with complete (cd1) or partial (cd2) direct instruction are considerably higher compared to complete self-regulated learning (cd4). At the beginning of the course, when students do not know about the demands of the statistics course, the odds in favour of the designs with complete or partial direct instruction (cd1, cd2) are about 6 and 5.3 times higher compared to complete self-regulated learning design of cd4 ($\pi_{11}/\pi_{41} = 0.405/0.068 = 6.0$ and

$\pi_{21}/\pi_{41} = 0.357/0.068 = 5.3$). Before the first test, the odds in favour of the designs cd1 and cd2 decrease, but are still about 2.4 and 2.6 times higher compared to cd4. There are similar odds at the second and third time points. These effects are considerable. They mean that students overwhelmingly state that they need the presence of an instructor in basic statistic courses in a ratio of about five to one. The effect declines as the course progresses, but the effect is still substantial even after the first statistics test.

In summary, our model provides a simple and easy to use method for exploring longitudinal paired comparisons in the presence of attrition. Additionally, the parameters are readily interpretable by practitioners. It is hoped that this paper will encourage the future use of longitudinal paired comparisons as a research design for exploring changes in attitude.

Acknowledgements We would like to thank Walter Katzenbeisser (WU Vienna) for helpful comments and David Fletcher (Otago) for fruitful e-mail discussions on the overdispersed multinomial. We also thank the referees for helping us to improve our paper. Originally we planned to write this paper with Reinhold Hatzinger (WU Vienna) but he sadly died before we started writing. We acknowledge the inspiration of Reinhold Hatzinger to this work on longitudinal paired comparisons and would like to thank him for his encouragement and support, and for fruitful discussions.

References

Agresti A (2013) Categorical Data Analysis 3rd Edition. Wiley, New York

Aitkin M, Francis B, Hinde J, Darnell R (2009) Statistical Modelling in R. Oxford Statistical Science Series 35, Oxford University Press, Oxford

Böckenholt, U and Dillon, W (1997) Some new methods for an old problem: modeling preference changes and competitive market structures in pretest market data. J Marketing Res 34(1):130–142

Bradley R, Terry M (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39(3/4):324–345

Cattelan M, Varin C, Firth D (2013) Dynamic Bradley-Terry modelling of sports tournaments. J R Stat Soc Ser C 62(1):135–150

Dittrich R, Hatzinger R, Katzenbeisser W (1998) Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. J R Stat Soc Ser C 47(4):511–525

Dittrich R, Hatzinger R, Katzenbeisser W (2002) Modelling dependencies in paired comparison data. A log-linear approach. Comput Stat Data Anal 40:39–57

Dittrich R, Hatzinger R, Katzenbeisser W (2004) A log-linear approach for modelling ordinal paired comparison data on motives to start a PhD programme. Stat Model 4:181–193

Fahrmeir L, Tutz G (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. J Am Stat Assoc 89(428):1438–1449

Francis B, Dittrich R, Hatzinger R, Penn R (2002) Analysing partial ranks by using smoothed paired comparison methods: an investigation of value orientation in Europe. J R Stat Soc Ser C 51(3):319–336

Francis B, Dittrich R, Hatzinger R, Humphreys L (2014) A mixture model for longitudinal partially ranked data. Commun Stat Theor M 43(4):722–734

Glickman M (1993) Paired comparison models with time-varying parameters. PhD dissertation, Harvard University Department of Statistics, Cambridge USA

Glickman M (1999) Parameter estimation in large dynamic paired comparison experiments. J R Stat Soc Ser C 48(3):377–394

Glickman M (2001) Dynamic paired comparison models with stochastic variances. J Appl Stat 28(6):673–689

Grand A, Greimel-Fuhrmann B, Hatzinger R (2013) Analyse der Präferenzen von Studierenden für verschiedene Lehrveranstaltungsdesigns - Ein loglineares Modell für wiederholte Präferenzmessungen anhand von Paarvergleichen [Analysis of students' preferences for different course designs – A log-linear model for repeated preference measurements on paired comparisons]. Empir Pädag 27(2):183–205

Hatzinger R (2012) prefmod: Utilities to fit paired comparison models for preferences. R package version 0.8-31. http://cran.R-project.org/web/packages/prefmod/index.html

Jackson CH (2011) Multi-state models for panel data: The msm package for R. Journal of Statistical Software 38(8):1–29, URL http://www.jstatsoft.org/v38/i08/

Kousgaard N (1984) Analysis of a sound field experiment by a model of paired comparisons with explanatory variables. Scand J Stat 11:51–57

Lindsey JK (1992) The analysis of stochastic processes using GLIM. Lecture Notes in Statistics 72, Springer-Verlag, Berlin

Lindsey JK (2004) Statistical analysis of stochastic processes in time. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge

Matthews J, Morris K (1995) An application of Bradley-Terry-type models to the measurement of pain. J R Stat Soc Ser C 44(2):243–255

R Development Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org

Shen C, Chen Y (2012) Model selection for generalized estimating equations accommodating dropout missingness. Biometrics 68(4):1046–54

Sinclair C (1982) GLIM for preference. In: Gilchrist R (ed) GLIM 82, Proceedings of the international conference on generalised linear models. Springer Lecture Notes in Statistics, 14 , Springer-Verlag, Berlin, pp 164–178

Turner H, Firth D (2012) gnm: Generalized nonlinear models. R package version 1.0-6. http://CRAN.R-project.org/package=gnm

Ver Hoef J, Boveng P (2007) Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 28(11):2766–2772