Optimal design of paired comparison experiments in the presence of within-pair order effects

Peter Goos

Universiteit Antwerpen

Heiko Großmann

Queen Mary, University of London

This paper presents a systematic approach to dealing with within-pair order effects in paired comparison experiments, which have historically received substantial attention in many application areas, including psychology, transportation, marketing and medicine. It describes how optimal designs can be constructed for paired comparison experiments involving several attributes. It extends earlier work on design of experiments in the presence of within-pair order effects for a single qualitative attribute and fills a void in the recent design literature on choice experiments with multiple attributes.

Keywords: Bradley-Terry model, choice-based conjoint experiments, conditional logit model, *D*-optimal design, two-alternative forced-choice experiments, within-pair order effects

1 Introduction

Paired comparison studies are utilized in various fields of study, including marketing, transportation, environmental and health economics, psychology and sensometrics. This is because performing several pairwise comparisons of products or services may prove to be a more realistic approach to constructing an overall ranking than ranking a large set of alternatives directly (see, e.g., Agresti 2002). In paired comparison studies, sets of two alternatives are offered to respondents, who then have to indicate the alternative they like most. In some paired comparison studies, the respondents are asked to state how strong their preference is. The alternatives offered in each of the pairs may be real or hypothetical products or services (in marketing experiments), travel modes (in transportation studies), and health states (in health economics), but they can also be medical treatments (in the context of pain measurement), verbal crime descriptions (in psychology and law) or cue combinations (in psychophysics). In certain application areas, the alternatives are referred to as profiles and they are described in terms of properties or attributes. The attributes can be qualitative or quantitative. The purpose of the paired comparison experiment is to assess the importance of each of the attributes to the respondents and to determine the weights the respondents attach to the different levels of the attributes. In the marketing literature, these weights are sometimes called part-worths. In psychophysics, paired comparison experiments are often named two-alternative forced-choice experiments, and the alternatives are referred to as stimuli, which consist of one or more signals or cues. Undoubtedly, this terminology is new to readers who are not familiar with the most recently published applications of paired comparison studies. As a matter of fact, paired comparison studies have historically been used mainly to compare levels of a single qualitative attribute.

The design of paired comparison studies involving multiple attributes has received considerable attention recently. The work published in this domain typically borrows ideas from a research area called optimal design of experiments, which seeks experimental designs that allow the most efficient estimation or prediction of some unknown quantities. As one of the main goals of paired comparison studies is to determine the part-worths, the optimal design approach is appropriate to design them. Applications of optimal design to paired comparison experiments can be found in Offen and Littell (1987), van Berkum (1987, 1989), El-Helbawy et al. (1994), Graßhoff et al. (2003, 2004), Sándor and Wedel (2001), Kessels et al. (2006, 2009), Street and Burgess (2004, 2007), and Großmann et al. (2009). It should be pointed out that some of the publications in this enumeration do not restrict attention to paired comparison experiments only, but deal with choice experiments in general (where respondents evaluate sets of two or more profiles).

Remarkably, none of the listed references takes into account the potential existence of within-pair order effects, even though this topic has received a substantial amount of attention in the literature. Early discussions and references on the importance of the order in which to present pairs and the order in which to present alternatives within pairs in the method of paired comparisons involving one qualitative attribute can be found in Ross (1934, 1939) and Wherry (1938). The work by Ross is considered definitive by David (1963). Some technical follow-up work on Ross's designs can be found in Simmons and Davis (1975) and Cloete et al. (1988). The designs proposed by Ross possess the following properties: (i) an alternative is compared with every other alternative; (ii) the order of the pairs is such that the time gap between two consecutive presentations of the same alternative is maximal; and (iii) every alternative appears as often on the left as on the right of a pair.

The incorporation of within-pair order effects in the analysis of data from paired comparison studies received explicit attention by Scheffé (1952), who suggested an analysis of variance approach for preferences expressed on a 7- or 9-point scale and who found a highly significant order effect in a taste experiment. Beaver and Gokhale (1975) point out that in paired comparison studies involving psychophysical stimuli, the effect of the order of presentation may be more important than the magnitudes of the stimuli themselves. Therefore, they propose a modification to the Bradley-Terry model (Bradley & Terry 1952) involving one additive order effect per ordered pair, and find evidence of the presence of such order effects in a weights-judging experiment (see also Beaver 1977). Davidson and Beaver (1977) suggest another modification to the Bradley-Terry model to cope with within-pair order effects. Rather than additive order effects, they suggest using multiplicative ones. They argue that multiplicative order effects arise naturally from the setting of the linear model and offer several technical advantages over the model with additive order effects. Augustin (2004) provides further support for the superiority of the model involving multiplicative order effects, and Fienberg (1979) re-examines the multiplicative order effects model from a technical viewpoint. Harris (1957) presented a constant additive bias model, which is similar to that of Davidson and Beaver (1977) except that it is based on the Thurstone (1927) model (see, e.g., Critchlow and Fligner 1991). Davidson and Beaver (1977) find evidence of order effects in the weights-judging experiment mentioned above, as well as in an experiment involving different food mixes. Applications of the multiplicative order effects model in wine tasting and psychophysics can be found in Lukas (1991) and Wickelmaier and Choisel (2006), respectively. In each of these applications, evidence of order effects was found. Such evidence was also found by van der Waerden et al. (2006) in choice experiments with three alternatives per choice set for studying transport mode decisions, and by Chrzan (1994) in choice experiments when investigating brand preferences in marketing. Matthews and Morris (1995) took into account a potential order effect in a paired comparison experiment for the measurement of pain, but did not obtain convincing evidence of its existence.

A point worth stressing is that, while all the models proposed for incorporating within-pair order effects in the statistical analysis of paired comparison data allow for different order effects for every pair of alternatives, the applications described in Davidson and Beaver (1977), Matthews and Morris (1995) and Wickelmaier and Choisel (2006), use a single order effect only. The authors report a satisfactory model fit, despite the model simplification. Similarly, van der Waerden et al. (2006) use a model with order effects that are common for every choice set. In this paper, we will therefore also focus on the situation where one order effect is common to all the pairs presented in the paired comparison experiment.

Even though the literature on the design of paired comparison experiments involving multiple attributes has not paid any attention to within-pair order effects, the profiles within pairs usually have to be ordered, either spatially or temporally. Obvious examples are taste experiments, experiments in which respondents watch commercials and studies in which respondents try out the alternatives. Also, in experiments where profiles are described verbally or displayed using pictures, there is always one alternative that has to come first. It is therefore prudent to design experiments so that the experimental results are not distorted by potential order effects. One approach often taken by researchers is to randomize the order of the alternatives within each pair. An even better approach is to design the paired comparison experiment so that the parameter estimates are estimated independently from the order effect. Unlike the randomization-based one, this approach uses a systematic ordering of the alternatives within every pair. Below are two simple examples of experiments involving several attributes in which order effects are likely to be present.

Example 1. A race bicycle constructor is interested in testing several new configurations. Several test riders are available during the months of July and August for testing the bicycles. Each rider tries two bicycles: one in July and one in August. The three attributes under study are the type of frame (two levels: classic or sloping frame), the type of wheels (three levels: Campagnolo Hyperon, Mavic Ksyrium SL, Shimano WH-7701) and the groupset (two levels: Campagnolo Record, Shimano Dura-Ace).

Example 2. A global player on the beer market recently introduced 1.5 liter PET bottles in Russia. The bottle, made of an improved polyethyleneteraphtalate (PET), guaranteed a 90-day shelf life of the product. In order to find the best possible composition of the bottle, several experiments were performed. In addition to experiments focusing on the chemical and physical properties of the bottled beer (e.g. carbon dioxide, proportion of sulphite, and turbity), taste experiments were performed. Among the factors investigated in the experiments were the proportion of nylon (two levels: 2%, 4%), the amount of iron (Fe) in the PET (1500 ppm, 2500 ppm), the type of cap (two levels) and, in some of the tests, the time elapsed between the bottling and the consumption of the beer (two levels: 45 days, 90 days).

The situation where the alternatives are obtained by combining levels of different attributes, or factors, is referred to as a factorial treatment structure or a situation with structured treatments in the experimental design literature. When the alternatives are just levels of a single qualitative variable, as in most of the literature on within-pair order effects, the literature on experimental design uses the term unstructured treatments. In this paper, the focus is on factorial treatment structures. This complicates the problem of designing paired comparison experiments. This is due to the fact that the experimenter has to decide which combinations of attribute levels to use in the experiment out of many possible ones, on top of determining which pairs to use and in which order to present the alternatives within each pair. Note that the order of the pairs themselves (which was considered by Ross (1934)) is much less important under factorial treatment structures than in the case of unstructured treatments because of the large number of different alternatives which arise from combining levels of several attributes.

2 Models for paired comparison studies

In the recent literature on paired comparison experiments, two different models can be found for the data analysis. The first model is the conditional logit model or the Bradley-Terry model, which is used when the respondents are asked to choose between the two alternatives in every pair. The pairs are then called choice sets. The second model is a linear model, which is used whenever the respondents have to indicate their preference for one of the two alternatives in a pair on a continuous scale.

2.1 Pairwise choice experiments

A popular statistical model used for analyzing data from paired comparison experiments is the logit model for paired evaluations proposed originally by Zermelo (1929) as an analytically convenient alternative to the probit model originally proposed by Thurstone (1927). The logit model was popularised by Bradley and Terry (1952) and embedded in a regression framework by McFadden (1974). A review of the early literature on the background concerning the approach, its use, applications and extensions can be found in Bradley (1976).

The Bradley-Terry model supposes that the probability π_{12i} of preferring alternative 1 over alternative 2 in the ith paired comparison can be expressed as

$$\pi_{12i} = \frac{\exp(u_{1i})}{\exp(u_{1i}) + \exp(u_{2i})}$$

where u_{1i} and u_{2i} represent the utilities attached to the two alternatives in paired comparison i. Hence the probability $\pi_{21i} = 1 - \pi_{12i}$ that alternative 2 is preferred over alternative 1 can be written as $\pi_{21i} = \exp(u_{2i})/(\exp(u_{1i}) + \exp(u_{2i}))$. If follows that for each pair i the choice probabilities as well as their ratio depend only on the utility difference $u_{1i} - u_{2i}$. In situations where the alternatives are described by means of several attributes, the utilities are modelled using the linear predictor

$$u_{ji} = \mathbf{x}'_{ji}\boldsymbol{\beta}, \quad j = 1, 2,$$

where $\mathbf{x}_{ji}$ is the vector containing the coded levels of the attributes of the jth alternative in the ith paired comparison. The utility difference $u_{1i} - u_{2i} = (\mathbf{x}_{1i} - \mathbf{x}_{2i})'\boldsymbol{\beta}$ is then a linear function of the unknown parameter vector $\boldsymbol{\beta}$. A maximum likelihood estimate of $\boldsymbol{\beta}$ can be obtained, for example, by means of any software routine for simple logistic regression using the components of the vectors $(\mathbf{x}_{1i} - \mathbf{x}_{2i})'$ as the predictors for the ith pair.

2.2 Linear paired comparison studies

Another type of paired comparison study, which in particular covers the model of Scheffé (1952), are linear paired comparison experiments. In this type of study, rather than just choosing the alternative they prefer the respondents have to indicate on a scale how strong their preference is. One quantitative response is then observed per paired comparison which again depends on the difference vector $(\mathbf{x}_{1i} - \mathbf{x}_{2i})'$. More precisely, the response is described by the linear model

$$Y_i = u_{1i} - u_{2i} + \varepsilon_i = (\mathbf{x}_{1i} - \mathbf{x}_{2i})'\boldsymbol{\beta} + \varepsilon_i,$$

where the ε_i for different pairs i are typically assumed to be independent with zero mean and variance σ^2.

2.3 Order effects

In situations where the alternatives to be compared cannot be presented simultaneously, order effects might influence the outcome of the comparison. Also, even when the two alternatives are simultaneously presented on a screen, an order effect might be present as one of the alternatives would typically be read or looked at first.

The simplest way to model the presence of an order effect is to assume it is additive and modify, say, the expression $u_{1i} = \mathbf{x}'_{1i}\boldsymbol{\beta}$ for the utility of the first alternative in each pair by adding a parameter δ. The relevant utility differences in both the Bradley-Terry and the linear paired comparison model then become

$$\delta + (\mathbf{x}_{1i} - \mathbf{x}_{2i})'\boldsymbol{\beta},$$

where δ represents the order effect which here is assumed to be the same for all pairs.

3 Information matrix and design optimality criteria

A precise estimation of the two models outlined above requires carefully designed paired comparisons. The quality of a design for paired comparison experiments is expressed using the Fisher information matrix on the unknown model parameters in the model. The criterion most often used for selecting designs for paired comparison studies is the D-optimality criterion, which seeks designs that maximize the determinant of the Fisher information matrix.

The Fisher information matrix on the parameter vector $\boldsymbol{\beta}$ in the conditional logit model in Section 2.1 is equal to

$$\mathbf{I}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{\pi_{12i}\mathbf{x}_{1i}\mathbf{x}'_{1i} + \pi_{21i}\mathbf{x}_{2i}\mathbf{x}'_{2i} - (\pi_{12i}\mathbf{x}_{1i} + \pi_{21i}\mathbf{x}_{2i})(\pi_{12i}\mathbf{x}_{1i} + \pi_{21i}\mathbf{x}_{2i})'\}, \tag{1}$$

where N represents the total number of paired comparisons in the experiment. In order to compare designs with possibly different numbers of pairs N, the normalized Fisher information matrix $\mathbf{M}(\boldsymbol{\beta}) = N^{-1}\mathbf{I}(\boldsymbol{\beta})$, which summarizes the information per pair, is usually used.

Both $\mathbf{I}(\boldsymbol{\beta})$ and $\mathbf{M}(\boldsymbol{\beta})$ depend on the unknown parameter vector $\boldsymbol{\beta}$ through the probabilities π_{12i} and π_{21i}. This is inconvenient because it implies that the D-optimal paired comparison design depends on the unknown model parameters. One way to deal with this is to adopt a Bayesian approach which takes into account any available prior information. However if, as in the works of Street et al. (2001) and Street and Burgess (2004, 2007), it is assumed that the probabilities for the two alternatives in each choice set are equal to 1/2 or equivalently that $\boldsymbol{\beta} = \mathbf{0}$, then the normalized Fisher information matrix can be written as

$$\mathbf{M}(\boldsymbol{\beta}) = \frac{1}{4N}\sum_{i=1}^{N}(\mathbf{x}_{1i} - \mathbf{x}_{2i})(\mathbf{x}_{1i} - \mathbf{x}_{2i})'. \tag{2}$$

Note that the right-hand side of this equation does not depend on $\boldsymbol{\beta}$. The assumption of equal probabilities is sometimes realistic because it reflects the fact that the researcher is ignorant concerning the part-worths of the attributes under investigation. In matrix notation, the right-hand side of (2) can be re-written as

$$\frac{1}{4N}\mathbf{X}'\mathbf{X}, \tag{3}$$

where for each paired comparison i the matrix $\mathbf{X}$ contains a row given by the difference vector $(\mathbf{x}_{1i} - \mathbf{x}_{2i})'$. The matrix $\mathbf{M} = N^{-1}\mathbf{X}'\mathbf{X}$ can be recognized as the normalized information matrix in the linear paired comparison model in Section 2.2 (Graßhoff et al. (2004)). Since that matrix is proportional to the matrix in (3) it follows that if $\boldsymbol{\beta}$ is assumed to be a zero vector, then the same designs are D-optimal for the conditional logit and the linear paired comparison model.

In either model, including an order effect δ can be regarded as adding a two-level factor to the set of predictor variables or attributes. The corresponding difference vector for pair i then has an additional component and can be written as $(c, \mathbf{x}_{1i} - \mathbf{x}_{2i})'$, where c is a constant that depends on the coding but which is the same for all pairs. Gathering the row vectors for all pairs into a single matrix we obtain the matrix $(c\mathbf{1}_N, \mathbf{X})$ where $\mathbf{1}_N$ is a column vector of length N with all elements equal to 1 and $\mathbf{X}$ is defined as before.

Denoting the normalized Fisher information matrix in the conditional logit model including an order effect by $\mathbf{M}(\delta, \boldsymbol{\beta})$, it follows that, for $\delta = 0$ and $\boldsymbol{\beta} = \mathbf{0}$,

$$\mathbf{M}(\delta, \boldsymbol{\beta}) = \frac{1}{4N}(c\mathbf{1}_N, \mathbf{X})'(c\mathbf{1}_N, \mathbf{X}) = \frac{1}{4}\begin{pmatrix} c^2 & \frac{c}{N}\mathbf{1}_N'\mathbf{X} \\ \frac{c}{N}\mathbf{X}'\mathbf{1}_N & \mathbf{M} \end{pmatrix}, \tag{4}$$

where $\mathbf{M}$ is defined as before. Moreover, up to a constant factor the matrix on the right-hand side of (4) coincides with the information matrix in the extended linear paired comparison model including the order effect.

4 Optimal designs

If each column of the matrix $\mathbf{X}$ sums to zero it follows that $\mathbf{1}'_N\mathbf{X}$ is a row vector of zeros. Hence in this case, it follows from (4) that the effects of interest can be estimated independently of the order effect. The design should therefore be orthogonally blocked with respect to the two levels of the factor corresponding to the order effect.

Example 3. To illustrate how orthogonality of the order effect and the attributes can be achieved consider the design in Table 1 for a main-effects model, where the first attribute has two and the second attribute has three levels. For each attribute each level appears the same number of times in Alternative 1 and Alternative 2. Converting the attribute levels of Alternative 1 in each pair using effects-type coding results in a matrix $\mathbf{X}_1$ whose rows are the vectors previously denoted as $\mathbf{x}_{1i}$, $i = 1, \ldots, N$. Similarly, applying the coding to Alternative 2 in each of the pairs gives rise to a matrix $\mathbf{X}_2$ whose ith row is the vector $\mathbf{x}_{2i}$. It is then not difficult to verify that each column of $\mathbf{X} = \mathbf{X}_1 - \mathbf{X}_2$ sums to zero. Other common types of coding yield the same result.

Table 1: Paired comparison design with treatment effects orthogonal to order effect

Pairs				$\mathbf{X}_1$			$\mathbf{X}_2$			$\mathbf{X} = \mathbf{X}_1 - \mathbf{X}_2$		
Alternative 1		Alternative 2		Alternative 1			Alternative 2			Difference		
1	1	2	2	1	1	0	−1	0	1	2	1	−1
1	2	2	3	1	0	1	−1	−1	−1	2	1	2
1	3	2	1	1	−1	−1	−1	1	0	2	−2	−1
2	1	1	2	−1	1	0	1	0	1	−2	1	−1
2	2	1	3	−1	0	1	1	−1	−1	−2	1	2
2	3	1	1	−1	−1	−1	1	1	0	−2	−2	−1

In addition to having orthogonality between the attributes and the order effects it is important that the paired comparison study provides maximum information. Most criteria for measuring the information content of a design are based on the normalized information matrix. The most commonly used of these is the D-criterion which aims to maximize the determinant of $\mathbf{M}(\delta, \boldsymbol{\beta})$.

Under the hypothesis of equal choice probabilities or equivalently $\boldsymbol{\beta} = \mathbf{0}$, and assuming a main-effects model with K factors at l_k levels, $k = 1, \ldots, K$, which are coded as in Table 1, it follows from the results of Graßhoff et al. (2004) that a design with normalized information matrix equal to

$$\mathbf{M}(\boldsymbol{\beta}) = \frac{1}{4}\begin{bmatrix} \mathbf{M}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{M}_K \end{bmatrix} \tag{5}$$

is D-optimal for the model without an order effect, where for every k

$$\mathbf{M}_k = \frac{2}{l_k - 1}\begin{bmatrix} 2 & 1 & \dots & 1 \\ 1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 2 \end{bmatrix}.$$

Note that $\mathbf{M}_k$ has $l_k - 1$ rows and columns. Moreover, if for such a design $\mathbf{1}'_N\mathbf{X}$ is a zero row vector, then the design is also D-optimal in the model containing the order effect. More precisely, under the hypothesis of equal choice probabilities the design is D-optimal for estimating both δ and $\boldsymbol{\beta}$ and D-optimal for the subset of parameters contained in $\boldsymbol{\beta}$. In that case $\mathbf{M}(\delta, \boldsymbol{\beta}) = \mathrm{diag}[1, \mathbf{M}(\boldsymbol{\beta})]$.

Example 3 continued. Assuming that the choice probabilities in the model including the order effect are equal, the normalized information matrix of the design in Table 1 can be seen to be equal to

$$\mathbf{M}(\delta, \boldsymbol{\beta}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 1/4 \\ 0 & 0 & 1/4 & 1/2 \end{bmatrix},$$

which implies that the design is D-optimal.

In general, under the hypothesis of equal choice probabilities, a D-optimal design for estimating the main effects of K attributes with l_k levels, $k = 1, \dots, K$, in the model including the order effect can be constructed by adapting the methods described in Graßhoff et al. (2004). Below such a generalization is described for a technique based on orthogonal arrays (see, e.g., Hedayat et al. (1999)) which is particularly useful when the attributes have only two or three levels.

1. For each attribute k with an even number of levels l_k form all $s_k = l_k(l_k - 1)$ ordered pairs of levels.

2. Similarly, for each attribute k for which l_k is odd form all $s_k = l_k(l_k - 1)/2$ pairs of levels and arrange them such that each level appears in the first position as often as in the second position.

3. Find the smallest orthogonal array $OA(N; s_1, \dots, s_K; 2)$ of strength 2 with N rows and K columns.

4. Using the pairs from Steps 1 and 2 each row of the orthogonal array can be expanded into a pair of alternatives. This is done by replacing the s_k different symbols in the kth column of the orthogonal array as follows:

a) If attribute k has an even number of levels l_k, then replace all instances of a given symbol in column k of the orthogonal array with one of the s_k ordered pairs from Step 1. This is done in such a way that each of the s_k symbols is replaced by a different pair.

b) If attribute k has an odd number of levels l_k, then similarly replace every instance of a given symbol in the kth column with one of the s_k ordered pairs from Step 2.

5. The output of Step 4 is an $N \times K$ array of pairs of attribute levels which is then split into two $N \times K$ arrays $\mathbf{L}$ and $\mathbf{R}$. Array $\mathbf{L}$ is formed using the left element in each pair. Array $\mathbf{R}$ is formed using the right element in each pair.

6. The ith pair in the final design is then obtained by using the ith row of $\mathbf{L}$ to specify the first and the ith row of $\mathbf{R}$ to specify the second alternative.

Example 4. Table 2 illustrates the construction of a D-optimal design with 18 paired comparisons for a model with a single two-level and seven three-level attributes. For the two-level attribute there are $s_1 = 2$ ordered pairs $(1, 2)$ and $(2, 1)$ which are used to replace the symbols in the first column of the orthogonal array on the left of the table. For each of the remaining attributes we have $s_k = 3$, $k = 2, \ldots, 8$, and the symbols in the kth column of the orthogonal array are replaced by the pairs $(1, 2)$, $(2, 3)$ and $(3, 1)$. The central portion of the table shows the resulting array which is then further decomposed into the $N = 18$ pairs of the optimal design.

5 Algorithmic approach

The combinatorial approach outlined above relies on the existence of an orthogonal array of the type $OA(N; s_1, \ldots, s_K; 2)$. If all s_k values required in the design construction are equal to two or three, then many orthogonal arrays with small numbers of rows N exist. As s_k is two or three only for two- and three-level attributes, this implies that optimal paired comparison designs with a reasonable number of paired comparisons can be constructed for attributes with two or three levels. However, for four- and five-level attributes, the required values for s_k equal $l_k(l_k - 1) = 4(4 - 1) = 12$ and $l_k(l_k - 1)/2 = 5(5 - 1)/2 = 10$. If several s_k values with these magnitudes are required, then it is impossible to find orthogonal arrays with a manageable size N.

In such cases, a computerized-search algorithm can be utilized to seek an optimal paired comparison design of pre-specified size N. Computerized-search algorithms have the advantage that they can be used for any value of N, unlike the combinatorial approach based on orthogonal arrays. However, they suffer from two drawbacks. First, a computerized-search algorithm cannot guarantee that an optimal design will be found, nor that each of the model parameters can be estimated independently from the order effect. Second, the application of computerized-search algorithms may become problematic if the number of

Table 2: Construction of a D-optimal design with 18 paired comparisons for a model with one attribute at two and seven attributes at three levels

																Optimal design															
Orthogonal array								Array after replacing symbols								**L**								**R**							
0	0	0	0	0	0	0	0	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
0	0	1	2	2	0	1	1	(1,2)	(1,2)	(2,3)	(3,1)	(3,1)	(1,2)	(2,3)	(2,3)	1	1	2	3	3	1	2	2	2	2	3	1	1	2	3	3
0	0	2	1	2	1	0	2	(1,2)	(1,2)	(3,1)	(2,3)	(3,1)	(2,3)	(1,2)	(3,1)	1	1	3	2	3	2	1	3	2	2	1	3	1	3	2	1
1	0	1	1	0	2	2	0	(2,1)	(1,2)	(2,3)	(2,3)	(1,2)	(3,1)	(3,1)	(1,2)	2	1	2	2	1	3	3	1	1	2	3	3	2	1	1	2
1	0	2	0	1	2	1	1	(2,1)	(1,2)	(3,1)	(1,2)	(2,3)	(3,1)	(2,3)	(2,3)	2	1	3	1	2	3	2	2	1	2	1	2	3	1	3	3
1	0	0	2	1	1	2	2	(2,1)	(1,2)	(1,2)	(3,1)	(2,3)	(2,3)	(3,1)	(3,1)	2	1	1	3	2	2	3	3	1	2	2	1	3	3	1	1
0	1	1	1	1	1	1	0	(1,2)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)	(2,3)	(1,2)	1	2	2	2	2	2	2	1	2	3	3	3	3	3	3	2
0	1	2	0	0	1	2	1	(1,2)	(2,3)	(3,1)	(1,2)	(1,2)	(2,3)	(3,1)	(2,3)	1	2	3	1	1	2	3	2	2	3	1	2	2	3	1	3
0	1	0	2	0	2	1	2	(1,2)	(2,3)	(1,2)	(3,1)	(1,2)	(3,1)	(2,3)	(3,1)	1	2	1	3	1	3	2	3	2	3	2	1	2	1	3	1
1	1	2	2	1	0	0	0	(2,1)	(2,3)	(3,1)	(3,1)	(2,3)	(1,2)	(1,2)	(1,2)	2	2	3	3	2	1	1	1	1	3	1	1	3	2	2	2
1	1	0	1	2	0	2	1	(2,1)	(2,3)	(1,2)	(2,3)	(3,1)	(1,2)	(3,1)	(2,3)	2	2	1	2	3	1	3	2	1	3	2	3	1	2	1	3
1	1	1	0	2	2	0	2	(2,1)	(2,3)	(2,3)	(1,2)	(3,1)	(3,1)	(1,2)	(3,1)	2	2	2	1	3	3	1	3	1	3	3	2	1	1	2	1
0	2	2	2	2	2	2	0	(1,2)	(3,1)	(3,1)	(3,1)	(3,1)	(3,1)	(3,1)	(1,2)	1	3	3	3	3	3	3	1	2	1	1	1	1	1	1	2
0	2	0	1	1	2	0	1	(1,2)	(3,1)	(1,2)	(2,3)	(2,3)	(3,1)	(1,2)	(2,3)	1	3	1	2	2	3	1	2	2	1	2	3	3	1	2	3
0	2	1	0	1	0	2	2	(1,2)	(3,1)	(2,3)	(1,2)	(2,3)	(1,2)	(3,1)	(3,1)	1	3	2	1	2	1	3	3	2	1	3	2	3	2	1	1
1	2	0	0	2	1	1	0	(2,1)	(3,1)	(1,2)	(1,2)	(3,1)	(2,3)	(2,3)	(1,2)	2	3	1	1	3	2	2	1	1	1	2	2	1	3	3	2
1	2	1	2	0	1	0	1	(2,1)	(3,1)	(2,3)	(3,1)	(1,2)	(2,3)	(1,2)	(2,3)	2	3	2	3	1	2	1	2	1	1	3	1	2	3	2	3
1	2	2	1	0	0	1	2	(2,1)	(3,1)	(3,1)	(2,3)	(1,2)	(1,2)	(2,3)	(3,1)	2	3	3	2	1	1	2	3	1	1	1	3	2	2	3	1

attributes is large and/or the number of levels of several attributes is large.

These drawbacks can best be illustrated by considering an example. Suppose that a researcher wishes to construct a paired comparison design with $N = 36$ paired comparisons for studying eleven two-level attributes and twelve three-level attributes. An optimal design for this problem can be constructed combinatorially by means of the method in Section 4 using an orthogonal array for which $N = 36$, $s_1 = \cdots = s_{11} = 2$ and $s_{12} = \cdots = s_{23} = 3$. A design constructed in this way guarantees that all the model parameters can be estimated independently from the order effects. Also note that the design is saturated, that is the number of estimated parameters including the order effect is equal to the number of pairs.

An alternative approach would be to construct a design using a computerized-search algorithm for constructing optimal designs in blocks. This is because every pair in a paired comparison design can be viewed as a block of size two. The best known algorithms in the literature on the D-optimal design of blocked experiments for factorial treatment structures are those of Atkinson and Donev (1989), Cook and Nachtsheim (1989), Goos and Vandebroek (2001), and Goos and Donev (2006). A similar algorithm has been implemented in various software packages, for example in the SAS procedure OPTEX. However, because these algorithms require the construction of a candidate set with all possible alternatives, they cannot be used for the problem involving eleven two-level attributes, twelve three-level attributes and an order effect. The reason for this is that it is impossible to list the set of $2^{11}3^{12}$ alternatives. The only algorithm that is capable of handling that complex a problem is the coordinate-exchange algorithm of Meyer and Nachtsheim (1995), which does not require the list of all possible alternatives as an input and which has been implemented in JMP. One thousand runs of the coordinate-exchange algorithm in JMP required about 2 minutes and 40 seconds and produced a design which was clearly not optimal. Its performance in terms of the D-optimality criterion relative to the combinatorially constructed optimal design amounts to 95.21%. With 10,000 tries of the coordinate-exchange algorithm, the efficiency of the computer-generated design relative to the combinatorially constructed one is 95.34%. Most, but not all, of the model parameters can be estimated independently from the order effects. Excel files containing the designs discussed in this section are available for download at the authors' web pages.

From this example, it is clear that the computerized-search algorithm is able to produce a reasonably good solution for a challenging paired comparison design problem, but not the optimal design. The less challenging the nature of the design problem, the more likely a computerized-search algorithm will find the optimal design rather than just a nearly optimal design. For smaller design problems, the point-exchange algorithms can be used as alternatives to the coordinate-exchange algorithm.

6 Conclusion

The early literature on paired comparisons paid much attention to the problem of within-pair order effects, but the more recent work on the design of paired comparison studies and choice experiments completely ignores this potential cause of bias. We provide a detailed literature study and revise how within-pair order effects can be modelled and taken into account when constructing paired comparisons designs. We also describe a simple combinatorial construction method that guarantees optimal designs in many situations and that outperforms computerized-search algorithms for problems involving large numbers of attributes. Because of the use of paired comparisons in disciplines as diverse as marketing, transportation, environmental and health economics, psychology and sensometrics, the combinatorial construction method will be useful for a broad range of applications. For example, the optimal design with 36 paired comparisons for 11 two-level attributes and 12 three-level attributes discussed in Section 5 can be easily modified for situations with smaller numbers of two- and three-level attributes. More specifically, if there are a two-level attributes and b three-level attributes, this can be done by simply selecting any a two-level columns and any b three-level columns. Because of the similarity between paired comparison and two-color micro-array studies, the construction method can also be applied to such micro-array studies if a difference in fluorescence intensity is expected between the red and the green color.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edn, New York: Wiley.

Atkinson, A. C. & Donev, A. N. (1989). The construction of exact D-optimum experimental designs with application to blocking response surface designs, *Biometrika* **76**: 515–526.

Augustin, T. (2004). Bradley-Terry-Luce models to incorporate within-pair order effects: Representation and uniqueness theorems, *British Journal of Mathematical and Statistical Psychology* **57**: 281–294.

Beaver, R. J. (1977). Weighted least-squares analysis of several univariate Bradley-Terry models, *Journal of the American Statistical Association* **72**: 629–634.

Beaver, R. J. & Gokhale, D. V. (1975). A model to incorporate within-pair order effects in paired comparisons, *Communications in Statistics* **4**: 923–939.

Bradley, R. A. (1976). Science, statistics, and paired comparisons, *Biometrics* **32**: 213–232.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs I. The method of paired comparisons, *Biometrika* **39**: 324–345.

Chrzan, K. (1994). Three kinds of order effects in choice-based conjoint analysis, *Marketing Letters* **5**: 165–172.

Cloete, W. G., Cloete, I. & von Gadow, K. (1988). An algorithm for presenting pairs in optimum orders, *EDV in Medizin und Biologie* **19**: 75–77.

Cook, R. D. & Nachtsheim, C. J. (1989). Computer-aided blocking of factorial and response-surface designs, *Technometrics* **31**: 339–346.

Critchlow, D. E. & Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear-models, and their implementation on GLIM, *Psychometrika* **56**: 517–533.

David, H. A. (1963). *The Method of Paired Comparisons*, London: Charles Griffin.

Davidson, R. R. & Beaver, R. J. (1977). On extending the Bradley-Terry model to incorporate within-pair order effects, *Biometrics* **33**: 693–702.

El-Helbawy, A. T., Ahmed, E. A. & Alharbey, A. H. (1994). Optimal designs for asymmetrical factorial paired comparison experiments, *Communications in Statistics: Simulation* **23**: 663–681.

Fienberg, S. E. (1979). Log linear representation for paired comparison models with ties and within-pair order effects, *Biometrics* **35**: 479–481.

Goos, P. & Donev, A. N. (2006). Blocking response surface designs, *Computational Statistics and Data Analysis* **51**: 1075–1088.

Goos, P. & Vandebroek, M. (2001). D-optimal response surface designs in the presence of random block effects, *Computational Statistics and Data Analysis* **37**: 433–453.

Graßhoff, U., Großmann, H., Holling, H. & Schwabe, R. (2003). Optimal paired comparison designs for first-order interactions, *Statistics* **37**: 373–386.

Graßhoff, U., Großmann, H., Holling, H. & Schwabe, R. (2004). Optimal designs for main effects in linear paired comparison models, *Journal of Statistical Planning and Inference* **126**: 361–376.

Großmann, H., Graßhoff, U. & Schwabe, R. (2009). Approximate and exact optimal designs for paired comparisons of partial profiles when there are two groups of factors, *Journal of Statistical Planning and Inference* **139**: 1171–1179.

Harris, W. P. (1957). A revised law of comparative judgment, *Psychometrika* **22**: 189–198.

Hedayat, A. S., Sloane, N. J. A. & Stufken, J. (1999). *Orthogonal Arrays. Theory and Applications*, Springer, New York.

Kessels, R., Goos, P. & Vandebroek, M. (2006). Comparing algorithms and criteria for designing Bayesian conjoint choice experiments, *Journal of Marketing Research* **43**: 409–419.

Kessels, R., Jones, B., Goos, P. & Vandebroek, M. (2009). An efficient algorithm for constructing Bayesian optimal choice designs, *Journal of Business and Economic Statistics* **27**: 279–291.

Lukas, J. (1991). BTL-Skalierung verschiedener Geschmacksqualitäten von Sekt, *Zeitschrift für Experimentelle und Angewandte Psychologie* **38**: 605–619.

Matthews, J. N. S. & Morris, K. P. (1995). An application of Bradley–Terry-type models to the measurement of pain, *Applied Statistics* **44**: 243–255.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, *in* P. Zarembka (ed.), *Frontiers in Econometrics*, New York: Academic Press, pp. 105–142.

Meyer, R. K. & Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs, *Technometrics* **37**: 60–69.

Offen, W. W. & Littell, R. C. (1987). Design of paired comparison experiments when the treatments are levels of a single quantitative variable, *Journal of Statistical Planning and Inference* **15**: 331–346.

Ross, R. T. (1934). Optimum orders for the presentation of pairs in the method of paired comparisons, *The Journal of Educational Psychology* **25**: 375–382.

Ross, R. T. (1939). Optimal orders in the method of paired comparisons, *Journal of Experimental Psychology* **25**: 414–424.

Sándor, Z. & Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs, *Journal of Marketing Research* **38**: 430–444.

Scheffé (1952). An analysis of variance for paired comparisons, *Journal of the American Statistical Association* **47**: 381–400.

Simmons, G. J. & Davis, J. A. (1975). Pair designs, *Communications in Statistics* **4**: 255–272.

Street, D. J., Bunch, D. S. & Moore, B. (2001). Optimal designs for 2^k paired comparison experiments, *Communications in Statistics: Theory and Methods* **30**: 2149–2171.

Street, D. J. & Burgess, L. (2004). Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments, *Journal of Statistical Planning and Inference* **118**: 185–199.

Street, D. J. & Burgess, L. (2007). *The construction of optimal stated choice experiments: Theory and methods*, Wiley, Hoboken, NJ.

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

van Berkum, E. E. M. (1987). Optimal paired comparison designs for factorial and quadratic model, *Journal of Statistical Planning and Inference* **15**: 265–278.

van Berkum, E. E. M. (1989). Reduction of the number of pairs in paired comparison designs and exact designs for quadratic models, *Computational Statistics and Data Analysis* **8**: 93–107.

van der Waerden, P., Borgers, A., Timmermans, H. & Bérénos, M. (2006). Order effects in stated-choice experiments, *Transportation Research Record: Journal of the Transportation Research Board* **1985**: 12–18.

Wherry, R. J. (1938). Orders for the presentation of pairs in the method of paired comparisons, *Journal of Experimental Psychology* **23**: 651–660.

Wickelmaier, F. & Choisel, S. (2006). Modeling within-pair order effects in paired-comparison judgments, *in* D. E. Kornbrot, R. M. Msetfi & A. W. MacRae (eds), *Fechner Day 2006. Proceedings of the 22nd Annual Meeting of the International Society for Psychophysics*, The ISP, St. Albans, UK, pp. 89–94.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximum-Problem der Wahrscheinlichkeitsrechnung, *Math. Zeit.* **29**: 436–460.