MODELING WITHIN-PAIR ORDER EFFECTS IN PAIRED-COMPARISON JUDGMENTS

Florian Wickelmaier[1,2] and Sylvain Choisel[1,3]
[1]*Sound Quality Research Unit, Dept. of Acoustics, Aalborg University, Denmark*
[2]*Dept. of Psychiatry, Ludwig-Maximilians-University of Munich, Germany*
[3]*Bang & Olufsen A/S, Struer, Denmark*
florian.wickelmaier@med.uni-muenchen.de; sc@acoustics.aau.dk

Abstract

Probabilistic choice models which generalize the Bradley-Terry-Luce (BTL) model (Luce, 1959) and the elimination-by-aspects (EBA) model (Tversky, 1972) are presented. These models account for the effect of presentation order within a pair of stimuli, and thereby allow for quantifying judgmental biases for the first or second presentation interval. Data were collected in an experiment where 39 subjects made pairwise choices between short musical excerpts reproduced in eight different audio formats (mono, stereo, and multichannel). Choice criteria were overall preference, and eight specific (spatial and timbral) auditory attributes. The results indicate that biases are well accounted for by the generalized modeling, and that for seven of the nine attributes, including preference, they significantly favored the second presentation interval.

Time and order effects in perception and decision making are phenomena met with great research interest by psychophysicists (see Hellström, 1985, for a review). In paired-comparison experiments stimuli are often presented sequentially. When subjects are asked to choose one of two stimuli with respect to a given criterion, their judgments might be influenced by the presentation order within that pair. As a consequence, there might be a perceptual or judgmental bias favoring either the first or the second presentation interval.

In the following section probabilistic choice models that account for the effect of presentation order within a pair are briefly introduced. Subsequently, an application is presented, where subjects were asked to judge upon overall preference and specific auditory sensations in an experiment on perceptual evaluation of reproduced sound. The models introduced were employed to quantifying within-pair order effects in the paired-comparison judgments.

Probabilistic choice models without order effect

Among the most widely used choice models for paired comparisons is the *Bradley-Terry-Luce (BTL)* model (Bradley & Terry, 1952; Luce, 1959). The BTL model predicts the probability, P_{xy}, of choosing stimulus x over stimulus y by

$$P_{xy} = \frac{u(x)}{u(x) + u(y)}, \tag{1}$$

where $u(\cdot)$ is a ratio scale (Luce, 1959) representing the strength or the weight of the stimuli. The BTL model requires the so-called *independence of irrelevant alternatives (IIA)* which in essence demands that the pairwise choices be made independently of the context introduced by a given pair. The IIA assumption has been criticized both on theoretical (e. g., Debreu, 1960) and empirical (e. g., Rumelhart & Greeno, 1971; Zimmer et al., 2004) grounds. In order to relax this assumption, Tversky (1972) introduced the *elimination-by-aspects (EBA)* model. According to EBA, each stimulus is characterized by a set of features or aspects. When choosing between two alternatives, only those aspects which the two alternatives do not have in common influence the decision. The probability of choosing x over y is then defined as

$$P_{xy} = \frac{\sum\limits_{\alpha \in x' \setminus y'} u(\alpha)}{\sum\limits_{\alpha \in x' \setminus y'} u(\alpha) + \sum\limits_{\beta \in y' \setminus x'} u(\beta)}, \tag{2}$$

where x' (y') denotes the set of aspects belonging to stimulus x (y), and $u(\cdot)$ is a ratio scale of the aspects. The set of aspects, x does not share with y, is indicated by $x' \setminus y'$.

Probabilistic choice models with order effect

Neither BTL nor EBA model take into account the presentation order of the stimuli within a pair, and that there might be a bias for the first or the second presentation interval. Therefore, Davidson & Beaver (1977) extended the BTL model and suggested a multiplicative order effect, $\vartheta_{xy} \geq 0$, that depends only on a given pair $\{x, y\}$. Let $P_{xy|x}$ denote the probability that x is chosen over y, given that x was presented first. Then the choice probabilities for the ordered pair (x, y) are defined as

$$P_{xy|x} = \frac{u(x)}{u(x) + \vartheta_{xy} \cdot u(y)} \qquad P_{xy|y} = \frac{\vartheta_{xy} \cdot u(x)}{\vartheta_{xy} \cdot u(x) + u(y)}, \tag{3}$$

where $u(\cdot)$ is a ratio scale and ϑ_{xy} is unique (Augustin, 2004). If ϑ_{xy} is smaller (greater) than one, the bias favors the first (second) choice interval; there is no order effect if $\vartheta_{xy} = 1$. A special case arises when $\vartheta_{xy} = \vartheta$, where the order effect is constant for all pairs.

In a similar manner, the EBA model may be extended to account for judgmental bias by introducing a multiplicative order effect. Choice probabilities are given by

$$P_{xy|x} = \frac{\sum\limits_{\alpha \in x' \setminus y'} u(\alpha)}{\sum\limits_{\alpha \in x' \setminus y'} u(\alpha) + \vartheta_{xy} \cdot \sum\limits_{\beta \in y' \setminus x'} u(\beta)}, \tag{4}$$

where parameters are defined as in Eq. 2 and Eq. 3. The probability $P_{xy|y}$ may be specified accordingly. Note that this generalized EBA model includes both the BTL model and the Davidson-Beaver model as special cases.

Method

Thirty-nine subjects took part in an experiment on the perceived quality of different audio formats (mono, stereo, and multichannel). Figure 1 displays the playback setup placed in an acoustically treated insulated room. Using this setup, four musical excerpts (*program materials:* two pop, two classical) of 5 s duration were played back to the listeners in eight different formats (*reproduction modes*), summarized in Table 1. Details of the stimulus rendering and presentation can be found in Choisel & Wickelmaier (in press).

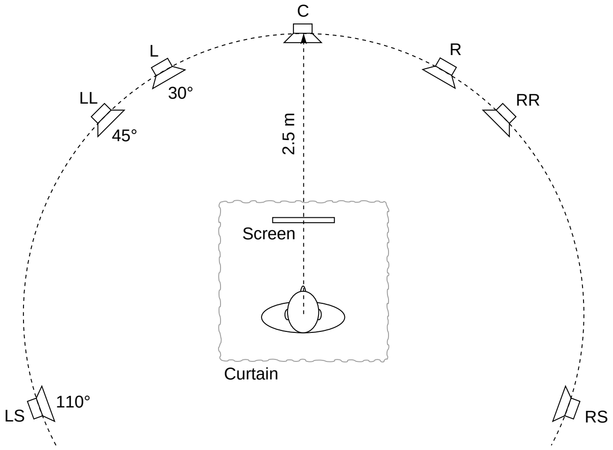

Figure 1: Playback setup consisting of seven loudspeakers: left (L), right (R), center (C), left-of-left (LL), right-of-right (RR), left surround (LS) and right surround (RS). This setup was symmetrically placed with respect to the width of the room and was hidden from the subject by an acoustically transparent curtain. A computer flat screen was used as a response interface.

For each pair of reproduction modes, the subjects were asked (in Danish) "Which of the two sounds is more..." followed by one of the following adjectives: *wide (bred), elevated (høj oppe), spacious (rummelig), enveloping (omsluttende), far ahead (langt foran), bright (lys), clear (tydelig)* and *natural (naturlig).* Two buttons on a computer screen, labeled A and B, were visually emphasized in turn (by changing their size) during playback to indicate which sound was playing. The response was given by clicking the button corresponding to the chosen sound. Each pair was judged only once. A completely balanced design (David, 1988, chapter 5) was employed to ensure that each stimulus occurred equally often in both presentation intervals. In addition, the within-pair order was balanced across subjects by having half of the subjects receiving the pairs in reverse order. The between-pair order was randomized for each subject.

Each participant gave 28 judgments per program material and auditory attribute in an experimental block. For a given attribute, all four program materials were completed sequentially in about 25 minutes. The order of the attributes and program materials was balanced across subjects using a Graeco-Latin square design. Overall pref-

Table 1: The eight reproduction modes used in the experiment: full name and loudspeakers active during playback (see Figure 1).

Name	Speakers
mono	C
phantom mono	L,R
stereo	L,R
wide stereo	LL,RR
matrix upmixing	L,R,LS,RS
Dolby Pro Logic II	L,R,C,LS,RS
DTS Neo:6	L,R,C,LS,RS
original 5.0	L,R,C,LS,RS

erence judgments were collected in a similar fashion, except that each subject received each pair in both orders, resulting in 56 preference judgments per musical excerpt and attribute.

The choice data collected in each presentation order were aggregated over subjects. Analysis of the choice frequencies thus obtained proceeded as follows: First, the two presentation orders were collapsed and the resulting data were modeled according to BTL (Eq. 1). Where BTL showed significant ($\alpha = .05$) lack of fit, the EBA model was employed (Eq. 2). Second, order-effect models were applied to the data observed in the two presentation orders; according to the previous results, either the Davidson-Beaver model (Eq. 3) or the EBA order-effect model (Eq. 4) were applied. Parameter estimation and model testing was performed using software described in Wickelmaier & Schmid (2004).

Results and discussion

Generally, the BTL model was found to fit the choice data well. For two attributes (*width* and *envelopment*) in one of the musical excerpts (Steely Dan), however, BTL had to be rejected; this indicates that IIA was violated and that more complex choice mechanisms might have played a role. In these two cases, EBA models (with nine aspect parameters each) accounted for the collected judgments sufficiently well (results not shown).

Table 2 shows the outcome of fitting order-effect models to the choice data. In columns three to five, the deviance statistic is reported as a lack-of-fit measure. Across all attributes and excerpts, the fit was fair to good apart from few exceptions (for lack of space, Table 2 displays the results for only four of the nine attributes). Columns six to eight show point and interval estimates of the order parameter, which was restricted to be equal for all pairs of a given attribute/excerpt data set. Estimates for ϑ ranged from 0.81 to 2.19 (mean 1.41). To illustrate, an estimate of $\hat{\vartheta} = 2$ would indicate that if two audio formats are equal with respect to a given attribute, the probability of choosing the second sound is *two times greater* than choosing the first sound.

For seven of the nine attributes, order effect parameters were significantly greater than one. This indicates that for the majority of attributes (including preference) there was a bias favoring the second presentation interval. For *brightness*, the order effect was not significant. Only for *distance*, order effect parameters were less than one, indicating that only for this attribute there is a tendency towards an advantage of the first interval.

Although the goodness of fit of the order effect models was generally found to be satisfactory, occasional misfits were observed. This might indicate that the assumption of a constant order effect across all pairs, $\vartheta_{xy} = \vartheta$, is too restrictive, and that the magnitude or even the direction of the bias varies with the pair. Further analysis to that effect is currently ongoing.

In the present study, the first sound played was always associated with the left button on the computer screen, and the second sound with the right button. Thus, the temporal effect of presentation order is confounded with a potential visual (or spatial) bias favoring the left or right button. Therefore, the current findings should be confirmed in further experiments where temporal presentation order and assignment to left/right button are counterbalanced.

In summary, the within-pair order effects encountered in this study were systematic and rather large. Across a variety of auditory attributes, judgments tended to be biased towards the second choice interval. The EBA order-effect model presented al-

Table 2: Goodness-of-fit statistics and parameter estimates for the order-effect models. Reported are the results of likelihood ratio tests of the Davidson-Beaver model (where df = 48) and the EBA order-effect model (where df = 47) vs. a saturated binomial model. Order effects (ϑ) and 95% confidence intervals have been estimated by maximum likelihood.

Attribute	Excerpt	χ^2	df	p	$\hat{\vartheta}$	2.5%	97.5%
Preference	Beethoven	59.80	48	0.118	1.55	1.38	1.72
–“–	Rachmaninov	82.44	48	0.001	1.86	1.66	2.06
–“–	Steely Dan	48.76	48	0.442	1.41	1.27	1.56
–“–	Sting	67.40	48	0.034	1.26	1.14	1.38
Envelopment	Beethoven	56.71	48	0.182	1.80	1.51	2.09
–“–	Rachmaninov	73.89	48	0.010	1.73	1.47	1.99
–“–	Steely Dan	63.38	47	0.056	1.36	1.16	1.57
–“–	Sting	39.99	48	0.788	1.15	0.99	1.31
Spaciousness	Beethoven	61.82	48	0.087	2.07	1.74	2.39
–“–	Rachmaninov	61.63	48	0.089	2.19	1.86	2.53
–“–	Steely Dan	61.89	48	0.086	1.27	1.08	1.46
–“–	Sting	63.06	48	0.071	1.48	1.24	1.72
Distance	Beethoven	57.92	48	0.155	0.83	0.73	0.93
–“–	Rachmaninov	41.57	48	0.732	0.89	0.78	1.00
–“–	Steely Dan	41.88	48	0.720	0.81	0.72	0.91
–“–	Sting	42.47	48	0.698	0.93	0.81	1.05

lows for quantification of such order effects, when context independence of the paired comparison judgments cannot be readily assumed.

Acknowledgements

This work was supported by the Centercontract on Sound Quality, Aalborg University.

References

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.

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

Choisel, S. & Wickelmaier, F. (in press). Evaluation of multichannel reproduced sound: scaling auditory attributes underlying listener preference. *Journal of the Acoustical Society of America.*

David, H. A. (1988). *The Method of Paired Comparisons.* New York: Oxford University Press.

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

Debreu, G. (1960). Review of R. D. Luce's Individual choice behavior: A theoretical analysis. *American Economic Review*, *50*, 186–188.

Hellström, Å. (1985). The time-order error and its relatives: mirrors of cognitive processes in comparing. *Psychological Bulletin*, *97*, 35–61.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Rumelhart, D. L. & Greeno, J. G. (1971). Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology*, *8*, 370–381.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*, 281–299.

Wickelmaier, F. & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, *36*, 29–40.

Zimmer, K., Ellermeier, W., & Schmid, C. (2004). Using probabilistic choice models to investigate auditory unpleasantness. *Acta Acustica united with Acustica*, *90*, 1019–1028.