Presentation-order effects for aesthetic stimulus preference

Mats P. Englund · Åke Hellström

Published online: 4 July 2012
© Psychonomic Society, Inc. 2012

Abstract For preference comparisons of paired successive musical excerpts, Koh (American Journal of Psychology, 80, 171–185, 1967) found time-order effects (TOEs) that correlated negatively with stimulus valence—the first (vs. the second) of two unpleasant (vs. two pleasant) excerpts tended to be preferred. We present three experiments designed to investigate whether valence-level-dependent order effects for aesthetic preference (a) can be accounted for using Hellström’s (e.g., Journal of Experimental Psychology: Human Perception and Performance, 5, 460–477, 1979) sensation-weighting (SW) model, (b) can be generalized to successive and to simultaneous visual stimuli, and (c) vary, in accordance with the stimulus weighting, with interstimulus interval (ISI; for successive stimuli) or stimulus duration (for simultaneous stimuli). Participants compared paired successive jingles (Exp. 1), successive color patterns (Exp. 2), and simultaneous color patterns (Exp. 3), selecting the preferred stimulus. The results were well described by the SW model, which provided a better fit than did two extended versions of the Bradley–Terry–Luce model. Experiments 1 and 2 revealed higher weights for the second stimulus than for the first, and negatively valence-level-dependent TOEs. In Experiment 3, there was no laterality effect on the stimulus weighting and no valence-level-dependent space-order effects (SOEs). In terms of the SW model, the valence-level-dependent TOEs can be explained as a consequence of differential stimulus weighting in combination with stimulus valence varying from low to high, and the absence of valence-level-dependent SOEs as a consequence of the absence of differential weighting. For successive stimuli, there were no important

effects of ISI on weightings and TOEs, and, for simultaneous stimuli, duration had only a small effect on the weighting.

Keywords Aesthetic preference · Presentation order · Time-order errors · Space-order errors · Visual perception · Audition · Math modeling

Fechner (1860) was the first to notice the ubiquitous and enigmatic systematic errors that occur in comparisons of successive and simultaneous stimuli, which make two physically equal stimuli subjectively different when compared: the *Zeitfehler* (time-order error) and the *Raumfehler* (space-order error). The time-order effect (TOE) and space-order effect (SOE) were defined as positive (vs. negative) for overestimation (vs. underestimation) of the first or left stimulus, respectively, relative to the second or right stimulus. Fechner (1876) also introduced experimental aesthetics, with scaling of aesthetic appreciation. However, he never combined those two subjects, which we attempt to do in the present article.

Since Fechner’s (1860) discovery, TOEs have been found for a wide range of modalities, including heaviness, tone loudness, line length, duration (see, e.g., Guilford, 1954; Hellström, 1985, for reviews), and brightness (Maeda, 1959). SOEs have been found in comparisons of, for example, line length (Hellström, 2003; Masin & Agostini, 1991) and brightness (Kellogg, 1931; Mattingley, Bradshaw, Nettleton, & Bradshaw, 1994). These presentation-order effects are, however, not restricted to comparisons of stimuli varied on a well-defined physical continuum. TOEs have also been found for *preference* judgments of visual stimuli (McLaughlin & Kermisch, 1997), auditory stimuli (Beebe-Center, 1932/1965; Koh, 1967; Koh & Hedlund, 1969), and odors (Beebe-Center, 1932/1965). For example, Koh (1967)

M. P. Englund · Å. Hellström (✉)
Department of Psychology, Stockholm University,
SE-106 91 Stockholm, Sweden
e-mail: hellst@psychology.su.se

investigated the possible existence of TOEs for musical pleasantness. She had participants rate the pleasantness of tape-recorded vocal excerpts (each lasting 60 s) and piano excerpts (each lasting 15 s) on a 9-step scale from *most pleasant* (1) to *most unpleasant* (9). Pairs of excerpts with equal mean ratings were then selected and presented successively, with an interstimulus interval (ISI) and an inter-trial interval (ITI) of about 6 and 10 s, respectively, to another sample of participants; these participants were to judge the direction and degree of the pleasantness difference between the excerpts in each pair, using a 7-step scale. For the vocal as well as the piano excerpts, large TOEs occurred that were highly correlated with the mean pleasantness rating: Participants consistently tended to prefer the second (a negative TOE) out of two pleasant excerpts and the first (a positive TOE) out of two unpleasant ones. On average, there was a slight tendency toward a negative TOE. Koh and Hedlund obtained similar results. Other experiments, reported by Beebe-Center (1932/1965), in which the pleasantness of auditory stimuli was compared, showed TOEs varying with the length of the ISI; TOEs were positive with an ISI of 1.5 s, but negative with ISIs of 2 s and longer. The results of these pleasantness comparisons resemble those of magnitude comparisons on traditional psychophysical continua—for instance, loudness (Hellström, 1979; Needham, 1935), heaviness (Hellström, 2000; Woodrow, 1933), and auditory and visual duration (Hellström, 2003). Analogous magnitude-dependent SOEs have been found in comparisons of line lengths (Hellström, 2003).

Although negative TOEs have been found more often than positive ones, a consistent finding has been that TOEs vary systematically with the intensity or magnitude level of the stimulation (see, e.g., Hellström, 1985). This was the basis for Hellström's (1979) *sensation-weighting (SW) model*. Hellström studied in detail the effects of stimulus magnitude on the TOE for loudness under different temporal stimulus presentation conditions. This led to the explanation of the TOE as a side effect of sensation weighting: The scaled subjective difference, d_{12} , between two compared stimuli is not the simple difference between their magnitudes, but can be modeled as the difference between two weighted compounds, one for each stimulus, where stimulus i ($i = 1, 2$) and a reference level (ReL) ψ_r are weighted by s_i and $(1 - s_i)$, respectively:

$$d_{12} = k\{[s_1\psi_1 + (1 - s_1)\psi_{r1}] - [s_2\psi_2 + (1 - s_2)\psi_{r2}]\} + b, \quad (1)$$

where k is a scale constant, ψ_1 and ψ_2 are the sensation magnitudes of the stimuli, and b is a term that accounts for effects apart from the weighting process (e.g., a possible response bias). The reason for the weighting-in of the ReLs is thought to be that information about average stimulus

magnitudes partially replaces the information about the specific stimulus magnitudes, in particular when this information is missing or noisy due to, for instance, memory loss (Hellström, 1985, 1989). Therefore, perceptual testing for changes in the difference between two stimuli by using the modified test variable defined by Eq. 1, instead of using the simple difference $k(\psi_1 - \psi_2)$, improves the discriminability of such changes (Hellström, 1985, 1989; Patching, Englund, & Hellström, *in press*) if the s values are optimized. A side effect, however, is the TOE or SOE, which can be defined, in subjective units, as the value of d_{12} in a pair of stimuli of equal magnitude (Hellström, 1985); using Eq. 1, setting $\psi_1 = \psi_2 = \psi$, and simplifying by assuming that $\psi_{r1} = \psi_{r2} = \psi_r$ yields

$$\text{TOE} = d_{12} = k(s_1 - s_2)(\psi - \psi_r) + b. \quad (2)$$

As mentioned above, in early research (typically with ISIs of at least a couple of seconds) negative TOEs were generally found, more so the higher was the pair's magnitude position in the series, and positive TOEs were found only for stimuli of low magnitudes (see, e.g., Woodrow, 1933). These results are explained in terms of Eq. 2 as a consequence of the weight relation $s_1 < s_2$ and the stimulus magnitudes in the pair being above the ReL ($\psi > \psi_r$; cf. Hellström, 2000). For brief stimuli and ISIs, Hellström (1979, 2003) found the opposite effect of stimulus magnitude and interpreted this as being due to the weight relation $s_1 > s_2$. For simultaneous line lengths, Hellström (2003) found the weight relation $s_{\text{left}} > s_{\text{right}}$, which may explain the finding that the SOEs were more positive for the longer lines.

Results showing that TOEs and SOEs vary with stimulus magnitude—in particular, those results showing changes in the signs of the TOE or SOE (e.g., Hellström, 2000, 2003; Koh, 1967)—seriously reduce the explanatory power of models that treat the TOE or SOE as a simple additive bias term (e.g., Beaver & Gokhale, 1975; Davidson & Beaver, 1977). Instead, the results provide evidence in favor of the SW model.

However, even though differential weighting of subjective stimulus magnitudes, along with Eq. 2, seems to offer an explanation (on the group level) to the results of Koh (1967; Koh & Hedlund, 1969), the presence of such weighting has not yet been investigated explicitly in individual comparisons of aesthetic stimuli. This weighting hypothesis suggests that preference comparisons are performed using judgment processes similar to the comparisons made in stimulus discrimination. Testing this weighting hypothesis thus promises to further the understanding of aesthetic comparison as well as that of stimulus comparison in general. Therefore, in the present study, we present three experiments designed to investigate the weighting hypothesis: in *Experiment 1*, via successive jingles; in *Experiment 2*, via

successive visual patterns; and in [Experiment 3](#), via simultaneous visual patterns. Specifically, in view of the results of Hellström (e.g., 1979, 2003), with different patterns of weighting and TOEs for different kinds of stimuli and large interindividual variability, we scaled the aesthetic values (valences) of the stimuli for each participant separately and investigated (a) whether TOEs analogous to those of Koh (1967) would be obtained with this within-subjects design and using briefly presented auditory stimulus sequences (jingles) and color patterns, with short ISIs; (b) whether or not the SW model can be used to account for the potential valence-level-dependent order effects and can offer a better fit than alternative models; (c) whether valence-level-dependent order effects (specifically, SOEs) for aesthetic preference also occur for color patterns with simultaneous presentation; and (d) whether the stimulus weighting and the order effects in aesthetic comparisons vary with ISI (successive stimuli) or duration (simultaneous stimuli).

General method

In each experiment, two samples of undergraduate psychology students participated to fulfill a course requirement. The participants in the first sample of each experiment took part in one experimental session, comprising from one to five comparison tasks with different kinds of stimuli (for the other tasks, see Hellström, 2003), and the second sample participated only in the three experiments presented here, all in one session. Participants made preference judgments of pairs consisting of successive jingles ([Exp. 1](#)) and of successive and simultaneous color patterns ([Exps. 2 and 3](#), respectively). The stimuli were presented in pairs on a Commodore Amiga 2000 computer with a Commodore 1081 color display screen in a quiet, softly lit room. The participant viewed the screen from a distance of approximately 45 cm. There were five stimuli in each experiment, each stimulus was paired with every other stimulus, and each pair was presented with four different ISIs ([Exps. 1 and 2](#)) or durations ([Exp. 3](#)). Thus, 80 stimulus pairs in total were presented in each experiment. After having read instructions presented on the screen, participants were offered an opportunity to ask questions regarding anything in the instructions that they might have felt was unclear. The participant started the experiment when ready. In each trial, the participants indicated the preferred stimulus by pressing a keyboard key—“1” for *first*, “2” for *second*, or “0” for *cannot decide*—and then finalizing the response by pressing Enter, before which point a correction could be made.

Difference scaling and data treatment

The scaling and model fitting were done individually for each participant. For each stimulus pair, the preference, or

subjective attractiveness difference, d_{12} was scaled by d^* , +100 for *first* (1), −100 for *second* (2), and 0 for *cannot decide* (0). Thus, the mean of d^* over the pairs was analogous to the $D\%$, or percent difference, measure, which indicates the difference between the percentages of *first-stimulus-greater* and *first-stimulus-less* responses in a set of stimulus pairs (Guilford, 1954, p. 306).

Equation 1 can be simplified to

$$d^* = B_1\psi_1 - B_2\psi_2 + C, \quad (3)$$

where $B_1 = ks_1$, $B_2 = ks_2$, and

$$C = k(\psi_{r1} - \psi_{r2} + s_2\psi_{r2} - s_1\psi_{r1}) + b. \quad (4)$$

In the employed scaling method, with m sets (ISIs or durations) and n stimuli, the estimated valence value, p^* (corresponding to ψ in [Eq. 3](#)) for each stimulus is obtained by scoring +100 for each choice of this stimulus, −100 for each choice of the other stimulus in a pair, and 0 for *cannot decide*, summing over the $2m(n-1)$ occurrences of a pair containing the stimulus (here, 32) and dividing by 200. Thus, for stimulus a :

$$p_a^* = \frac{1}{200} \sum_{k=1}^m \sum_{j=1, j \neq a}^n (d_{ajk}^* - d_{jak}^*), \quad (5)$$

where subscript j denotes the stimulus compared with a , and k is the set. Summing the n values of p_a^* , the terms in the numerator cancel out, so that the mean p^* value is 0. For a stimulus that is chosen every time and for one that is never chosen, p^* becomes $m(n-1)$ and $-m(n-1)$, respectively, so that in the present case the maximum and minimum possible values of p^* are +16 and −16. It should be emphasized that the scaling method is only based on counting preference choices and is independent of the choice of model to fit the resulting data.

In fitting the SW model for each participant and set (ISI or duration), a linear regression was computed across the 20 pairs with d^* for the pair as the dependent variable and the p^* values of the stimuli in the pair (representing their ψ values) as independent variables; thus,

$$d_{ijk}^* = B_{1k}p_i^* - B_{2k}p_j^* + C_k, \quad (6)$$

where subscripts i and j denote the compared stimuli, subscripts 1 and 2 their temporal or spatial (left and right, respectively) positions in the pair, and subscript k the set. As the mean of the n values of p^* is zero, for each set C_k equals the mean value of d_{ijk}^* .

There is a built-in restriction, which fixes the sum of the $2m$ values of B to $200/n$, so that, in the present experiments, the sum of the eight B values becomes 40. To see this, assume that each stimulus has, for the particular participant,

a true subjective valence value, ψ , which is invariant over sets. With the subscript “pred” denoting values predicted by the equation, for set k ,

$$d_{ijk, \text{pred}}^* = B_{1k}\psi_i - B_{2k}\psi_j + C_k \quad (7a)$$

and

$$d_{jik, \text{pred}}^* = B_{1k}\psi_j - B_{2k}\psi_i + C_k. \quad (7b)$$

Thus

$$(d_{ijk}^* - d_{jik}^*)_{\text{pred}} = (B_{1k} + B_{2k})(\psi_i - \psi_j). \quad (8)$$

For stimulus a , Eqs. 5 and 8 yield

$$(p_a^*)_{\text{pred}} = \frac{1}{200} \sum_{k=1}^m (B_{1k} + B_{2k}) \sum_{j=1, j \neq a}^n (\psi_a - \psi_j). \quad (9)$$

The last factor in Eq. 9, $\sum_{j=1, j \neq a}^n (\psi_a - \psi_j)$, is equal to $(n-1)\psi_a - \sum_{j=1, j \neq a}^n (\psi_j)$. Because the n values of ψ_j sum to 0, the sum of the $n-1$ values of $\psi_j, j \neq a$, becomes $-\psi_a$, so the last factor in Eq. 9 becomes $(n-1)\psi_a + \psi_a = n \cdot \psi_a$. This yields

$$(p_a^*)_{\text{pred}} = \frac{n}{200} \sum_{k=1}^m (B_{1k} + B_{2k}) \cdot \psi_a. \quad (10)$$

Thus, when $\sum_{k=1:m} (B_{1k} + B_{2k})$ is equal to $200/n$, $(p_a^*)_{\text{pred}}$ becomes equal to ψ_a . Equation 10 justifies the use of p^* , as defined here, as an estimate of ψ and fixes the sum of the eight B values to $200/n$ —here, 40.

The overall TOE in subjective units for each ISI or duration was computed as the mean scaled preference, across all stimulus pairs, of the first (left) stimulus over the second (right). This measure is equivalent to $D\%$, the difference between the percentages of “1” and “2” responses, and is termed $TOE\%$. The measure is also equal to C (Eq. 4), the d^* value predicted from Eq. 6 for a pair of stimuli with valence values equal to the mean valence—that is, zero.

Plots of the order effect against stimulus valence

The plots of the order effects (TOEs or SOEs) against the mean valences of the stimulus pairs were produced using the following procedure: For each participant, the stimulus pairs were ranked from the least to the most liked, using the rank order of the mean of their p^* values (described above). For each pair, a and b , the $TOE\%$ or $SOE\%$ value was calculated as the mean scaled preference, across the four ISIs or durations, of the first (left) stimulus over the second (right)—that is, $TOE\%(\text{or } SOE\%) = [\sum_{k=1:4} (d_{ab}^* + d_{ba}^*)]/8$. The TOEs

or SOEs and valence values for the stimulus pairs of corresponding rank orders were averaged across participants, and then the TOE or SOE values were plotted against the mean valence of the stimulus pairs.

Experiment 1: Successive jingles

It was predicted that, in accordance with the results of Hellström (1979, 1985, 2003), the SW model would yield a good fit, with differential weighting of the stimuli. It was further predicted that, in accordance with the results of Hellström (1979, 2003) for tone loudness, there would be a greater weight for the first stimulus than for the second for short ISIs, and vice versa for long ISIs. This would render the results for long ISIs similar to those of Koh (1967)—thus, with negative TOEs for well-liked stimuli and positive TOEs for disliked stimuli.

Method

Participants Two samples ($n_1 = 34$, $n_2 = 46$) participated to fulfill a course requirement. In total, 37 men and 43 women took part, all with normal hearing, and most of them psychology students—undergraduate (fulfilling a course requirement) or graduate (volunteering)—from the ages of 19–64 years ($M_{\text{age}} = 28.6$, $SD_{\text{age}} = 8.7$).

Stimuli and design The stimuli were five different jingles, sequences of seven sine-tone notes, played through the built-in loudspeaker at a comfortable level (measured at approximately 80 dBA from the position of the participant’s head). The tempered scale with $A_4 = 440$ Hz (subscript indicating the octave) was used to produce the jingles, which were (Jingle 1 [J1]) $D_5-C_5^\#-B_4-A_4-G_4-F_4^\#-E_4$; (J2) $E_5-F_4-F_4^\#-G_4-E_4-C_5-G_4$; (J3) $D_5-C_5-D_5-E_5-F_4-E_4-F_4$; (J4) $C_6-B_5-G_5-C_6-E_5-G_5-D_5$; and (J5) $F_6^\#-E_6-D_6-C_6^\#-B_5-A_5-B_5$. The notes within each jingle lasted 150 ms each and succeeded each other immediately. Thus, the duration of each jingle was 1,050 ms. The ISIs were 500, 1,000, 2,000, and 4,000 ms. Eighty pairs (one set for each ISI of the 20 pairs of different jingles, using both within-pair orders) were presented in a random order (different for each participant) with ISIs intermixed. The experiment, excluding instructions, lasted 12.1 min on average ($SD = 0.7$).

Results and discussion

As hypothesized, we found a valence-level dependent TOE, where the TOE correlated negatively with the valence level

of the stimuli (see Fig. 1);¹ the regression slope was significant, $t(9) = -6.99$, $p < .001$, but the intercept was not, $t(9) = -1.13$, $p = .292$. These results are analogous to those of Koh (1967), as well as to results found for comparisons of classic psychophysical stimuli (e.g., Guilford, 1954; Hellström, 1979, 1985, 2003). According to the SW model (e.g., Hellström, 1979, 2000), these results can be explained in terms of differential weighting of the stimuli, with a higher weight for the second stimulus (see Eq. 2). This explanation received support by analyses of the stimulus weighting. Specifically, estimates of B_1 and B_2 , which are proportional to the weights s_1 and s_2 , were obtained by regression, for each participant and ISI, of the preference ratings (d^*) on the valence values (p^*) of the compared stimuli. The mean across participants of the mean multiple R across ISIs was .705 ($SD = .141$, range .287–.930). The intraparticipant SD of the valences of the five stimuli, SD_{valence} , which can be seen as a measure of the consistency of preference judgments, and thus also of the degree to which the data lend themselves to modeling, had a mean value of 8.7 ($SD = 2.2$). The correlation across participants of SD_{valence} with the mean R across ISIs was .944.

Alternative modeling To our knowledge, no readily applicable models, other than the SW model, have been proposed that would be capable of accounting for the present results. Most other existing models of preference choice, such as the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) can be ruled out, because they are built on the assumption that stimulus comparison only involves a simple subtraction of the (transformed) stimulus magnitudes, and therefore cannot account for presentation-order effects. However, the BTL model was extended by Davidson and Beaver (1977) by including a parameter to account for the order effect: the multiplicative order-effect parameter γ_{ij} (with the presentation order i, j):

$$P(i > j|i, j) = \pi_i / (\pi_i + \gamma_{ij}\pi_j), \quad (11)$$

where π is the magnitude (here, valence) of the specific stimulus (i and j). This model will here be called the

¹ With the present scaling method, some of the stimulus pairs with neutral valence consisted of one highly liked and one highly disliked stimulus, respectively, and for these pairs, order effects were less likely. This may lead to a conservative estimate of the effect size of the valence-level-dependent TOE, but may also lead to a better fit of the regression line fitted to the data points in the plot. These two possibilities were tested by trimming the data by restricting the valence difference allowed between the stimuli within pairs. New analyses with these restrictions showed only marginal changes in the results; for example, allowing a valence difference of only 5 resulted in the fitted regression line $\hat{y} = -2.233x + 0.802$, $R = .871$. As for the nontrimmed data, the slope was significant, $t(9) = -5.01$, $p = .001$, but the intercept was not, $t(9) = 0.37$, $p = .720$.

Fig. 1 Experiment 1: Time-order effect (TOE%) plotted against mean stimulus valence. A fitted regression line is also displayed. A positive value of TOE% means a tendency to prefer the first stimulus

extended BTL (EBTL) model. An order effect is reflected by γ_{ij} deviating from 1, where $\gamma_{ij} > 1$ means an advantage for stimulus j and $\gamma_{ij} < 1$ a disadvantage for that stimulus (i.e., a negative and a positive TOE/SOE, respectively). As it stands, the EBTL model cannot account for the present results, because a given γ_{ij} value yields a positive *or* a negative order effect, which will not change sign with the stimulus magnitude (π). In particular, multiplying π_i and π_j by the same factor (i.e., changing the general valence level) will not change P . Replacing the P values by the corresponding log-odds ratios (logits), $\text{logit } P = \ln[P/(1 - P)]$, yields

$$\text{logit}[P(i > j|i, j)] = \ln(\pi_i/\pi_j) - \ln \gamma_{ij}. \quad (12)$$

As can be seen from Eq. 12, the probability of choosing one stimulus over the other is determined only by the ratio (π_i/π_j) , and $-\ln \gamma_{ij}$ merely enters as a constant added to $\text{logit } P$. As Englund and Hellström (2012b) remarked, “The only way for this kind of model to account for the present results is by letting the γ value change with the stimulus magnitude, and to the best of our knowledge, no one has suggested such an extension of the model” (p. 92). However, Davidson and Beaver (1977) did mention the possibility of letting γ_{ij} depend on the pair (i, j) .

In order to devise a BTL-type model that might challenge the SW model, we extended the EBTL model further, by making γ_{ij} dependent on the stimulus magnitude, and adapted it to the present type of data to create two alternative models, called EEBTL1 and EEBTL2. In EEBTL1, we let the γ_{ij} values differ between ISIs and be different for stimulus pairs with values of mean scaled valence below zero (non-preferred) and above zero (preferred), γ_{nonpref} and γ_{pref} , respectively. In addition, as the BTL models assume nonnegative

stimulus magnitudes, a constant, α , was added to the valence values separately for each ISI. This yielded $4 \times 3 = 12$ fitted parameters, one more than for the SW model, which has 11 (four C values and seven B values are fitted, as the eight B values sum to 40; see above). For each participant, the five stimulus scale values and the scaled response for each pair were entered. The responses were converted from $-100, 0$, and 100 to $0, 50$, and 100 , thus indicating a probability of preference for the first stimulus. The SW model was fitted to the data from each participant with SPSS 19 nonlinear regression (NLR), an iterative function-fitting program (yielding fitted parameter estimates equal or very close to those obtained by linear regression), and the EEBTL1 model was fitted with constrained nonlinear regression (CNLR), which differed from NLR by imposing constraints for the additive constant to ensure the nonnegativity of the resulting magnitudes. The goodness of fit of each model (here, for the entire data set for each participant) was expressed by NLR and CNLR in terms of the R^2 statistic [$= 1 - (\text{Residual sum of squares}/\text{Corrected sum of squares})$], the mean of which was .543 ($SD = .169$) for the SW, and .522 ($SD = .151$) for the EEBTL model. In 79 % of the cases, the SW model yielded a better fit, and a paired t test yielded $t(79) = 5.286$, $p < .001$. The advantage of the SW model increased with the goodness of fit of both models; $R^2_{\text{SW}} - R^2_{\text{EEBTL1}}$ correlated positively with R^2_{EEBTL1} ($r = .433$, $p < .001$). The correlations across participants of SD_{valence} with R^2_{SW} and R^2_{EEBTL1} were .926 and .919, respectively. SD_{valence} correlated positively with $R^2_{\text{SW}} - R^2_{\text{EEBTL1}}$ (.510, $p < .001$).

The EEBTL2 model differed from EEBTL1 by γ_{ij} being linearly dependent on the sum of the two scaled valence values [$\gamma_{ij} = \varepsilon + \beta(\psi_i + \psi_j)$], with separate values of ε , β , and α (see above) for each ISI. This makes 12 parameters, as for EEBTL1. For EEBTL2, the mean R^2 was .535 ($SD = .148$). In 65 % of the cases, the SW model yielded a better fit, and a paired t test on the R^2 values yielded $t(79) = 1.440$, $p = .154$. Again, the advantage of the SW model over EEBTL2 increased with the goodness of fit of the latter: $R^2_{\text{SW}} - R^2_{\text{EEBTL2}}$ correlated positively with R^2_{EEBTL2} ($r = .312$, $p = .005$). It also increased with SD_{valence} ($r = .515$, $p < .001$).

For the cases with better than average fit of EEBTL2 (i.e., $R^2 > .535$), the SW model (mean $R^2 = .673$, $SD = .092$) fit clearly better than EEBTL2 (mean $R^2 = .653$, $SD = .074$), $t(39) = 2.940$, $p = .005$. Comparing EEBTL1 and EEBTL2, the latter fit the data better, $t(79) = 2.701$, $p = .008$. $R^2_{\text{EEBTL2}} - R^2_{\text{EEBTL1}}$ correlated weakly negatively with R^2_{EEBTL1} ($r = -.228$, $p = .042$) and with SD_{valence} ($r = -.181$, $p = .108$).

None of the EEBTL models, with one more parameter than the SW model, could match the fit to the data of the latter. The less noisy the data (e.g., the greater the SD_{valence}), the better the fit of all models, but in particular of the SW model, and the clearer was the advantage of the SW model

over the EEBTL models, whereas there was no corresponding increase in advantage of EEBTL2 over EEBTL1. The parameters of the SW model are also more easily interpretable than those of EEBTL1 and EEBTL2. All of this speaks to the advantage of the SW model, which we selected for further analyses.

The B values from the SW model were submitted to a repeated measures ANOVA (multivariate approach), with Sample as between-subjects factor and Stimulus Position (first, second) and ISI as within-subjects factors. As predicted, the effect of position was significant: We found a higher mean weight for the second ($M = 5.80$, $SD = 1.25$) stimulus than for the first ($M = 4.20$, $SD = 1.25$), $F(1, 78) = 30.54$, $p < .001$, $\eta_p^2 = .281$. That is, $s_1 < s_2$, which indicates a greater impact of the second stimulus than of the first on the comparison, and is in accordance with typical findings from comparisons of stimuli presented successively (e.g., Hellström, 1979, 1985, 2000, 2003).

Further results from the ANOVA analysis showed that the main effect of ISI was nonsignificant, $F(3, 76) = 1.74$, $p = .165$, as was the interaction Position $\times$ ISI, $F(3, 76) = 1.27$, $p = .290$. These results are contrary to those from comparisons of loudness, in which Hellström (1979) found a higher weight for the first stimulus for short ISIs, but the opposite for long ones. One explanation may be that the jingles were easier to remember than the classic psychophysical stimuli. The Sample $\times$ ISI interaction approached significance, $F(3, 76) = 2.52$, $p = .064$, which was due to a significant interaction between the effect of sample and the linear effect of ISI, $t(76) = 2.28$, $p = .026$, $\eta_p^2 = .022$. However, as this effect concerned the average weights of the first and second stimuli, rather than their difference, it is not of particular theoretical or practical interest. Indeed, there were no significant effects of the interactions Sample $\times$ Position, $F(1, 78) = 0.72$, $p = .397$, and Sample $\times$ Position $\times$ ISI, $F(3, 76) = 1.11$, $p = .352$. The results, taken together, suggest that the weighting advantage of the second stimulus over the first in aesthetic comparison of jingles is a highly robust effect that is not affected easily by experimental manipulations. Given that jingles are stimuli stretched out in time, the higher weight for the second stimulus is likely due to memory decay of the stimulus presented first (e.g., Hellström, 1985); in SW theory, the stimulus weights are thought to be optimized to compensate for this memory decay by substituting lost information with average information of the stimulus series, which is reflected by lower weights (s values) in Eqs. 1 and 2. Accordingly, with greater memory decay of the first stimulus, participants may place a higher focus on the better-remembered second stimulus in the comparison, and thus compare the second stimulus to the first. Therefore, optimization of the weights and focusing on the better-remembered stimulus may be two sides of the same process. If so, the weighting difference should be indicative of the comparison direction (see, e.g., Englund & Hellström, 2012a, b).

The difference in the overall TOE% between the two samples was nonsignificant, $F(1, 78) = 0.50$, $p = .481$, and the mean TOE% was negative ($M = -1.09$, $SD = 16.83$) but not significantly different from zero, $t(79) = -0.58$, $p = .563$. The effect of ISI on TOE% was not significant, $F(3, 76) = 2.07$, $p = .111$, and the effect of ISI did not interact significantly with the effect of sample, $F(3, 76) = 0.38$, $p = .769$.

The reference level (ReL) in Eq. 2 was estimated using regression analysis. Specifically, simplifying Eq. 4 by assuming that $\psi_{r1} = \psi_{r2} = \psi_r$ yields

$$C = k(s_2 - s_1)\psi_r + b = (B_2 - B_1)\psi_r + b. \quad (13)$$

Then, using Eq. 13, ψ_r was estimated on roughly the same scale as the p^* values, as the slope in the regression, through the origin, of participants' individual means of C across ISIs on the corresponding means of $(B_2 - B_1)$; ψ_r was -0.714 ($SE = 0.632$, $p = .262$), which may be interpreted as being slightly below the average pattern in terms of valence. Including the intercept in this regression did not improve the fit significantly, $p = .974$, so it may be concluded that the model without the bias term b is adequate.

Experiment 2: Successive color patterns

Experiment 2 was designed to investigate whether the valence-level-dependent TOE found for jingles in **Experiment 1** and for musical excerpts by Koh (1967; Koh & Hedlund, 1969), the weighting pattern $s_1 < s_2$, and, hence, the SW explanation for the results, could be generalized to visual aesthetic stimuli. Therefore, the stimulus pairs in **Experiment 2** consisted of successive color patterns. It was hypothesized that the results would resemble those of **Experiment 1**, with a valence-level-dependent TOE (Koh, 1967; Koh & Hedlund, 1969), and thus would yield a greater weight for the second stimulus than for the first and no effect of ISI on the differential weighting of the stimuli.

Method

Participants Two samples ($n_1 = 33$, $n_2 = 46$) of undergraduate psychology students participated, 28 men and 51 women, from the ages of 19–50 years ($M_{\text{age}} = 26.6$, $SD_{\text{age}} = 6.7$).

Stimuli Rectangles of 70 (horizontal) $\times$ 100 (vertical) pixels (59×78 mm) were divided into four rectangles with two colors, A and B, in the pattern

A B
B A

The following five patterns (P1–P5) were used, defining A and B by the computer's 16 intensity levels (0–15) of, in

order, red, green, and blue; thus, the patterns, depicted here in the order (A) (B), were P1, (6 12 2) (8 14 10); P2, (13 15 9) (4 4 5); P3, (14 3 0) (1 14 1); P4, (5 0 15) (5 15 15); and P5, (12 4 14) (7 14 6). Including both within-pair orders, the stimulus combinations made up 20 different pairs in four different sets, one for each ISI (100, 300, 900, and 2,700 ms), yielding a total of 80 stimulus pairs. Each pattern was presented for 100 ms in the center of the screen, and the pairs from all of the sets were presented intermixed in a pseudorandom order that was the same for all participants in a sample, but differed between the samples (cf. Hellström, 2003, note 2).

Procedure The laboratory environment and the response mode were the same as in **Experiment 1**. The experiment, except for the instructions, lasted on average 8.9 min ($SD = 0.7$).

Results and discussion

Alternative modeling by the SW, EEBTL1, and EEBTL2 models was performed as in **Experiment 1**. The mean R^2 s were .675 ($SD = .120$) for SW, .637 ($SD = .109$) for EEBTL1, and .649 ($SD = .105$) for EEBTL2. In 95 % of the cases, SW fit better than EEBTL1, and a paired t test yielded $t(78) = 11.41$, $p < .001$. Also, in 86 % of the cases, SW fit better than EEBTL2, and a paired t test of the R^2 values yielded $t(78) = 7.61$, $p < .001$. $R^2_{\text{SW}} - R^2_{\text{EEBTL1}}$ correlated positively with R^2_{EEBTL1} , $r = .257$, $p = .022$, and $R^2_{\text{SW}} - R^2_{\text{EEBTL2}}$ correlated positively with R^2_{EEBTL2} , $r = .358$, $p = .0012$. The fit of EEBTL2 in terms of R^2 was significantly better than that of EEBTL1, $t(79) = 3.968$, $p < .001$. $R^2_{\text{EEBTL2}} - R^2_{\text{EEBTL1}}$ correlated negatively with R^2_{EEBTL1} ($r = -.250$, $p = .026$).

The intraparticipant SD_{valence} had a mean value of 10.9 ($SD = 1.6$). The correlation across participants of SD_{valence} with the mean R across ISIs was .940, and the correlations with the goodness of fit (R^2) for the SW, EEBTL1, and EEBTL2 models were, in order, .903, .887, and .878. SD_{valence} correlated positively with $R^2_{\text{SW}} - R^2_{\text{EEBTL1}}$ and with $R^2_{\text{SW}} - R^2_{\text{EEBTL2}}$ (.395, $p < .001$, and .512, $p < .001$, respectively) but not with $R^2_{\text{EEBTL2}} - R^2_{\text{EEBTL1}}$ ($-.156$, $p = .169$).

In accordance with expectations, we found a negative correlation between TOE and stimulus valence (see Fig. 2); the regression slope was significant, $t(9) = -4.11$, $p = .003$, but the intercept was not, $t(9) = -0.47$, $p = .648$.² These results are analogous to those of Koh (1967) and also to those from the jingle comparisons of **Experiment 1**, albeit with slightly lower effect sizes of the TOE% values. In order to investigate the

² As in **Experiment 1**, allowing a valence difference of 5 only changed the fitted regression line marginally ($\hat{y} = -1.869x - 2.074$, $R = .715$). As for the nontrimmed data, the slope was significant, $t(9) = -2.90$, $p = .020$, but the intercept was not, $t(9) = 0.60$, $p = .565$.

Fig. 2 Experiment 2: Time-order effect (TOE%) plotted against mean stimulus valence. A fitted regression line is also displayed. A positive value of TOE% means a tendency to prefer the first stimulus

stimulus weighting, the B values were estimated using the procedures of Experiment 1; the mean, across participants, of the mean multiple R s across ISIs was .812 ($SD = .093$, range = .304–.919).

The stimulus weighting was then analyzed by submitting the B values to an ANOVA for repeated measures with Within-Pair Stimulus Position (first, second) and ISI (100, 300, 900, 2,700 ms) as within-subjects factors.³ The ANOVA showed that the mean weight for the second stimulus ($M = 5.32$, $SD = 0.80$) was significantly higher than that for the first ($M = 4.68$, $SD = 0.80$), $F(1, 78) = 12.45$, $p < .001$, $\eta_p^2 = .138$. These results are in accordance with our hypotheses, with the results of Experiment 1, and also with previous research on comparisons of stimuli on physical continua (e.g., Hellström, 1979, 1985, 2000, 2003). Thus, these results demonstrate further the robustness of the weighting effect ($s_1 < s_2$) for successive presentation of aesthetic stimuli. The effect of ISI was not significant, $F(3, 76) = 0.97$, $p = .411$, nor was the interaction Position $\times$ ISI, $F(3, 76) = 1.21$, $p = .311$, which is also in accordance with the results of Experiment 1.

The average TOE% ($M = 0.40$, $SD = 8.77$) did not differ significantly from zero, $F(1, 78) = 0.16$, $p = .689$. An ANOVA for repeated measures, with Sample as a between-subjects factor and ISI as a within-subjects factor, showed that the main effect of ISI on TOE% was nonsignificant, $F(3, 75) = 1.52$, $p = .216$. There was a small difference in mean TOE%,

which approached significance, between the first sample ($M = -1.70$) and the second sample ($M = 1.90$), $F(1, 77) = 3.35$, $p = .071$. The marginally different overall TOE% values merely suggest a slight difference in the constant of the regression of TOE% on stimulus valence, but not in its slope. Indeed, a t test on the regression slopes of the two samples revealed no significant difference, $t(16) = -0.02$, $p = .767$. The interaction Sample $\times$ ISI approached significance, $F(3, 75) = 2.33$, $p = .081$. Analogously to the results of Experiment 1, this small difference in TOE% did not affect the slope in the regression of TOE% on stimulus valence—in a regression of TOE% on stimulus valence and its interaction with ISI, the interaction term was nonsignificant, $\beta = -.08$, $p = .504$.

Estimation of the ReLs for the two samples, using the procedures described in Experiment 1 (Eq. 13), showed no indication that the inclusion of a bias term and/or of two different ReLs would improve the model fit, which means that the simpler SW model (Eq. 2 with $b = 0$) is adequate to explain these data. The estimated ψ_s s for the first and second samples, respectively, were -0.99 ($SE = 0.93$, $p = .296$) and 2.06 ($SE = 0.675$, $p = .004$), and the difference between these two estimates was significant, $t(78) = 2.42$, $p = .018$. The negative value of ψ_r for the first sample represents a valence value lower than that for the average pattern, and the positive ψ_r for the second sample represents a valence value above that of the average pattern. The reason for this difference in ReLs between the samples is unknown, but whatever its reason, this difference and Eq. 2 explain the (nonsignificant) difference in the overall TOE%. More importantly, despite the difference in ReLs—and, hence, in TOE%—between the two samples, there were (as was checked by performing appropriate ANOVAs) no differences in the stimulus weighting, and therefore, no differences regarding the negative correlation of the TOE and stimulus valence. Taken together, these results suggest further that the stimulus weighting is an inherent part of the comparison process, where the net effect of this weighting ($s_1 < s_2$) can be described as an assimilation of the first pattern to the ReL (cf. Koh, 1967), which results in the valence-level dependence of the TOE.

Experiment 3: Simultaneous color patterns

The SOE does not seem to have been researched as extensively as has the TOE, at least in terms of the number of reports published on these topics. In the reports that we have found in the literature, there are some mixed results. Specifically, in typical psychophysical experiments, SOEs have been reported for comparisons of line length (Hellström, 2003; Masin & Agostini, 1991), of darkness (Kellogg, 1931) and of brightness (Mattingley et al., 1994). However, Patching et al. (in press) found only weak evidence of SOEs

³ Regarding the B values and their dependence on ISI and stimulus position, there were no significant differences between the two samples, as tested using repeated measures ANOVAs for a multivariate approach (Pillai tests). Therefore, the data sets for the two samples were collapsed for the analyses of the B values.

but more convincing evidence of TOEs in comparisons of brightness and of size. SOE-analogous effects (i.e., overestimation of stimuli in the left as compared to the right half of the visual field) have been found in comparisons of darkness, numerosity, and size (Nicholls, Bradshaw, & Mattingley, 1999; Rhode & Elias, 2007; Tant, Kuks, Kooijman, Cornelissen, & Brouwer, 2002). Similarly, in the literature regarding so-called *pseudoneglect* (i.e., the tendency to bisect lines noncentrally), it has been reported that healthy participants generally overestimate the left side of a prebisected line in forced choice tasks when judging on which side of center a line is bisected (Jewell & McCourt, 2000; McCourt, Freeman, Tahmahkera-Stevens, & Chaussee, 2001; McCourt & Garlinghouse, 2000; see also Rueckert, Deravanesian, Baboorian, Lacalamita, & Repplinger, 2002).

Regarding SOEs in preference comparisons of aesthetic stimuli, the literature seems even scarcer; we found only one study (Freimuth & Wapner, 1979). Freimuth and Wapner's participants made paired comparisons of paintings in which the pairs were composed of one painting and its mirror image. Freimuth and Wapner did not find SOEs in the paired comparisons of the mirrored paintings, but McLaughlin and Kermisch (1997) did find TOEs for similar stimuli. Analogous to these results are those of Patching et al. (*in press*), who found consistent evidence of TOEs but not of SOEs in comparisons of brightness and of size. Clearly, further investigation is needed regarding SOEs in preference comparisons of aesthetic stimuli, and preferably should be conducted in such a manner that the results of the SOE experiment can be compared directly with those of a matching experiment on TOEs (Exp. 2).

Therefore, in Experiment 3, we used the same stimuli as in Experiment 2 to investigate the effect of space order on aesthetic preferences for stimuli presented simultaneously and to test whether the data could be accounted for using the SW model. No strong a priori hypothesis regarding differential weighting could be made, because the previous results have been mixed. In studies on comparisons of line length, a greater weight for the left stimulus was found (Hellström, 2003; Masin & Agostini, 1991), but such effects were considerably weaker in the study by Patching et al. (*in press*). According to the SW model, an absence of differential weighting means an absence of SOEs, and vice versa, unless a response bias is involved. Therefore, the lack of SOEs in Freimuth and Wapner's (1979) study also suggested that we would find an absence of differential stimulus weighting.

Method

Participants Two samples ($n_1 = 36$, $n_2 = 46$) of undergraduate psychology students participated to fulfill a course requirement. Of the 82 who participated in the experiment,

23 were men and 59 were women, from the ages of 19–48 years ($M_{\text{age}} = 27.4$, $SD = 7.2$).

Apparatus, stimuli, and procedure The apparatus, stimuli, procedure, and preference scaling were the same as in Experiment 2, with the following exceptions: The stimuli were presented simultaneously, side by side, horizontally aligned in the middle of the screen, with a distance of 24 mm between the inner edges of each pair of color patterns, and the stimulus duration was varied instead of the ISI; the durations used were 100, 200, 400, and 800 ms. The experiment, excluding the instructions, lasted 12.4 min on average ($SD = 1.0$).

Results and discussion

Modeling by the SW model as well as the EEBTL1 and EEBTL2 models was performed as in Experiments 1 and 2. The mean R^2 s were .686 ($SD = .131$) for the SW, .652 ($SD = .114$) for the EEBTL1, and .656 ($SD = .111$) for the EEBTL2 models. In 89 % of the cases, the SW model yielded a better fit, and a paired t test yielded $t(81) = 9.74$, $p < .001$. In 82 % of the cases, SW fit better than EEBTL2, and a paired t test yielded $t(78) = 8.12$, $p < .001$. $R_{\text{SW}}^2 - R_{\text{EEBTL1}}^2$ correlated positively with R_{EEBTL1}^2 , $r = .453$, $p < .001$; $R_{\text{SW}}^2 - R_{\text{EEBTL2}}^2$ correlated positively with R_{EEBTL2}^2 , $r = .470$, $p < .001$. R_{EEBTL2}^2 was nonsignificantly higher than R_{EEBTL1}^2 , $t(79) = 1.250$, $p = .215$, and $R_{\text{EEBTL2}}^2 - R_{\text{EEBTL1}}^2$ correlated weakly negatively with R_{EEBTL1}^2 ($r = -.195$, $p = .080$).

The intraparticipant SD of the valences of the five stimuli had a mean value of 11.0 ($SD = 1.6$). The correlation across participants of the valence SD with the mean R across ISIs was .935, and with the goodness of fit (R^2) for the SW, EEBTL1, and EEBTL2 models were, in order, .904, .882, and .874. The valence SD correlated positively with $R_{\text{SW}}^2 - R_{\text{EEBTL1}}^2$ and with $R_{\text{SW}}^2 - R_{\text{EEBTL2}}^2$ (.574, $p < .001$, and .616, $p < .001$, respectively), but not with $R_{\text{EEBTL2}}^2 - R_{\text{EEBTL1}}^2$ (−.114, $p = .307$). Thus, again, a clear advantage of the SW model over the EEBTL models was found, and we may note that it occurred despite the lack of a strong average weighting asymmetry.

As can be seen in Fig. 3, there was no significant valence-level-dependent SOE—neither the regression slope nor the intercept was significant, $t(9) = -1.19$, $p = .267$, and $t(9) = -1.89$, $p = .096$,⁴ respectively—which is in line with the results of preference comparisons of paintings presented simultaneously (Freimuth & Wapner, 1979), but contrary to

⁴ As in Experiments 1 and 2, allowing a valence difference of only 5 did not change the results noteworthy ($\hat{y} = -0.618x + 3.109$, $R = .269$). After trimming the data, the slope was still nonsignificant, $t(9) = -0.79$, $p = .453$, and, although the intercept was more positive, it was still nonsignificantly different from zero, $t(9) = 0.69$, $p = .511$.

Fig. 3 Experiment 3: Space-order effect (SOE%) plotted against mean stimulus valence. A fitted regression line is also displayed. A positive value of SOE% means a tendency to prefer the left stimulus

those of comparisons of line length (Hellström, 2003). According to the SW model, the absence of a valence-level-dependent SOE suggests equal mean stimulus weights over durations ($M_{s1} \approx M_{s2}$; see Eq. 2). The analyses of the stimulus weighting provided support for this hypothesis.

The mean, across participants, of the mean multiple R s across stimulus durations was .816 ($SD = .097$, range .490–.946). Regression estimates of the B values were analyzed using a repeated measures ANOVA,⁵ with Within-Pair Stimulus Position (left, right) and Duration as within-subjects factors. There was no significant effect of stimulus position, $F(1, 81) = 0.67$. This suggests that there was no general differential weighting of the left ($M = 4.93$, $SD = 0.78$) and right ($M = 5.07$, $SD = 0.78$) stimuli, which is in contrast to previous results on comparisons of lines presented simultaneously (Hellström, 2003; Masin & Agostini, 1991). In the same ANOVA, the effect of the stimulus duration was also nonsignificant, $F(3, 79) = 1.16$, $p = .329$, but the Stimulus Position $\times$ Duration interaction approached significance, $F(3, 79) = 2.59$, $p = .059$; the interaction Stimulus Position $\times$ Cubic Effect of Duration was significant, $t(79) = -2.46$, $p = .016$, $\eta_p^2 = .024$ (see Fig. 4). Separate repeated measures ANOVAs for the B values for the left and the right stimuli, respectively, with

Fig. 4 Experiment 3: Mean B weights for the left (B_1) and right (B_2) stimuli, respectively, plotted against presentation duration

Stimulus Duration as a within-subjects factor, showed that the duration of the stimuli affected the weight for the left stimulus, $F(3, 79) = 3.76$, $p = .014$, $\eta_p^2 = .033$ [the cubic effect of duration was significant, $t(79) = 2.77$, $p = .007$, $\eta_p^2 = .031$], but not for the right stimulus, $F(3, 79) = 1.03$, $p = .383$.

The mean B weights plotted against the stimulus duration are displayed in Fig. 4. Paired-samples t tests showed that the differences between B_1 and B_2 for the respective durations were nonsignificant in all cases, $ps > .05$.

Using the same method as in Experiment 1, the mean of ψ_r was estimated to be -1.62 ($SE = 0.60$, $p = .009$), and there was no indication that the inclusion of a bias term and/or of two different ReLs (for the left and right stimuli, respectively) would improve the model fit. Hence, once again, the simpler form of the SW model (Eq. 2) is adequate to explain the present data. The average SOE% ($M = 0.93$, $SD = 0.88$) was not significantly different from zero, $F(1, 81) = 0.92$, and there was no significant effect of stimulus duration on the average SOE%, $F(3, 79) = 1.21$, $p = .310$. There were no differences in mean SOE% between the two samples, all $ps > .05$.⁶ In accordance with Eq. 2 without the bias term, the absence of a significant SOE in this experiment is explained by the absence of differential weighting, overall, of the compared stimuli ($s_1 \approx s_2$).

⁵ A regression of SOE% on stimulus valence with Stimulus Duration as a moderating factor showed no significant interaction between the stimulus valence and the duration, $\beta = -.123$, $p = .434$. This is in accordance with the SW model, as the differences between the weights for the left and right stimuli, respectively, were nonsignificant for all durations.

⁶ A regression of SOE% on stimulus valence with Stimulus Duration as moderating factor showed no significant interaction between the stimulus valence and the duration, $\beta = -.123$, $p = .434$. This is in accordance with the SW model, as the differences between the weights for the left and right stimuli, respectively, were nonsignificant for all durations.

General discussion

The main aims of the present study were to investigate (a) whether valence-level-dependent order effects, analogous to those reported by Koh (1967), would be obtained, and if so, whether the SW model could be used to account for these effects; (b) whether the valence-level-dependent order effects for aesthetic preference can be generalized to visual stimuli with successive and simultaneous presentation; and (c) to what extent the stimulus weighting and the order effects in aesthetic comparisons vary with ISI (for successive stimuli) or duration (for simultaneous stimuli). The present results were convincing regarding aim (a): For all of the experiments, there was an advantage in the fit of the data to the SW model in relation to the alternative models (EEBTL1 and EEBTL2) that increased with the fit of the inferior model and with the intraparticipant *SD* of the stimulus valences, scaled independently of models. Over the three experiments, the superiority of the SW model was clearer in Experiments 2 and 3 than in Experiment 1, where *SD*_{valence} was lower and all of the models fit somewhat less well. Still, in Experiment 1, results reminiscent of those of Koh (1967) were obtained using brief jingles rather than long musical excerpts and using shorter and varying ISIs, and the SW model could be used successfully to account for the valence-level-dependent TOEs, in terms of sensation weighting with a higher weight for the second stimulus than for the first.

Regarding aim (b), the results were not quite as conclusive. Whereas the valence-level-dependent TOEs found for the jingles (Exp. 1) were fully replicated for successive color patterns (Exp. 2), the results regarding analogous effects for simultaneous color patterns (Exp. 3) were mostly nonsignificant. With regard to the comparisons of aesthetic stimuli presented simultaneously in Experiment 3, there was no differential weighting, overall, which is contrary to reported results for comparisons of line length (Hellström, 2003; Masin & Agostini, 1991), but is partly in line with results for comparisons of brightness and for the size of light spots (Patching et al., *in press*). These discrepancies between research results may reflect differences between stimulus modalities. For example, differential stimulus weighting for simultaneous stimuli may be more pronounced for lines and perhaps for other geometric patterns.

Regarding aim (c), it was expected that, analogously with previous research on psychophysical comparisons (e.g., Hellström, 1979, 2000, 2003), the stimulus weighting and the order effects would vary with the length of the ISI (Exps. 1 and 2) or with duration (Exp. 3). However, there were no significant effects of ISI on the TOEs or SOEs. Regarding the stimulus weighting, we found a consistently higher weight for the second stimulus than for the first, and this effect did not vary with the length of the ISI, either, which is in contrast to previous findings in discriminations of

traditional psychophysical stimuli (e.g., Hellström, 1979, 2003). For example, Hellström (1979) found that, for tone loudness, the weight relation changed from $s_1 > s_2$ to $s_1 < s_2$ when ISIs increased beyond about 1 s. Hellström (e.g., 1985) suggested that the higher weight for the first stimulus, typically found for short ISIs with brief stimuli, is due to the unfinished processing of the first stimulus interfering with the processing of the second stimulus, and that the higher weight for the second stimulus than for the first, typically found for longer ISIs, is the result of greater memory loss for the first stimulus. In the present results, however, there was no such change in the weight relation as the length of the ISI varied from short to long. This difference in findings may be due to the aesthetic stimuli being easier to remember, or being processed more deeply, than are classic psychophysical stimuli, thus attenuating the effect of ISI.

An alternative, but not necessarily opposing, interpretation to that of memory loss is that the weight difference indicates the comparison direction (cf. Tversky, 1977). That is, the stimulus with the higher weight may be the *subject* that is compared to the other stimulus, the *referent*. This proposition has been made previously for preference comparisons of stimuli denoted by labels or written descriptions (e.g., Englund & Hellström, 2012a, 2012b; Houston, Sherman, & Baker, 1989; Wänke, Schwartz, & Noelle-Neumann, 1995) and has received support from explicit investigations by Englund and Hellström (2012a, 2012b; cf. Wänke, 1996). This idea seems congruent with Hellström's (e.g., 1985) suggestion that a lower weight magnitude reflects stimulus interference or adaptation to partial memory loss with the aim of optimizing stimulus discrimination. It seems reasonable to use the stimulus best represented in memory as the starting point of the comparison. For example, when the second stimulus has been presented after a relatively long ISI, it seems reasonable to try to compare it to the stimulus that was presented earlier. It may be noted that Hellström (1977, 1978) showed that making participants focus their responses on the first or the second of two tones, respectively, was the longer, the shorter, or the louder had no appreciable effects on the relative weightings of the two stimuli (where the second stimulus had a higher weight than the first). Thus, the effect of comparing the better (represented or) remembered stimulus to the worse on the comparison direction seems to be larger than the effect of changing the response instructions. Indeed, changing the response instructions does not necessarily change the comparison direction (e.g., Houston et al., 1989).

The results of the present study have demonstrated the continuity of preference comparisons with comparisons of physical magnitudes by showing valence-level-dependent order effects that are analogous to the magnitude-level-

dependent order effects for stimuli in classic psychophysical research (e.g., Guilford, 1954; Hellström, 1985). For example, earlier research (see Koh, 1967) showed TOEs for affective and aesthetic judgments that were either inconsistent or mainly negative. The latter type of result may have been due to successive stimulus presentation with the weighting pattern $s_1 < s_2$ and to stimuli of higher aesthetic value than their background, which is analogous to, for example, heaviness comparisons (Hellström, 2000). Indeed, this interpretation received ample support in the present study, due to the robustness, for successive presentation, of valence-level-dependent TOEs in which the TOE changed direction as stimuli changed from negative to positive valence, as expressed by the weight relation $s_1 < s_2$. Therefore, it seems that the common pattern of negative TOEs partly reflects psychophysicists' convenient use of stimuli above the background in the judged attribute, presented at a slow tempo. Models with assumptions of the comparison process being "simple subtraction"—that is, equal weighting of the compared stimuli, with or without a constant bias—cannot account for these data. Instead, the SW model, with differential stimulus weighting and a reference level, seems more adequate. However, how to interpret the stimulus weighting is not entirely understood. For discrimination of psychophysical stimuli, it has been suggested previously (e.g., Hellström, 1985) that low weights may reflect memory loss or processing interference, and for preference judgments of stimuli denoted by labels, it has been suggested that the weights indicate the comparison direction (Englund & Hellström, 2012a, 2012b). These two views may appear very different, but optimization of the weights and focusing on the better-remembered stimulus may be two sides of the same process. An understanding of how to interpret the stimulus weights should provide important information on the principles of stimulus comparison. Clearly, this is a natural topic for future research.

Author note This research was supported by the Swedish Research Council and by Stockholm University.

References

- Beaver, R. J., & Gokhale, D. V. (1975). A model to incorporate within-pair order effects in paired comparisons. *Communications in Statistics*, 4, 923–939.
- Beebe-Center, J. G. (1965). *The psychology of pleasantness and unpleasantness*. New York, NY: Russell & Russell (Original work published 1932).
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39, 324–345. doi:10.1093/biomet/39.3-4.324
- Davidson, R. R., & Beaver, R. J. (1977). On extending the Bradley–Terry model to incorporate within-pair order effects. *Biometrics*, 33, 693–702.
- Englund, M. P., & Hellström, Å. (2012a). *Beware how you compare: Comparison direction dictates stimulus-valence modulated presentation-order effects in preference judgment*. Submitted for publication.
- Englund, M. P., & Hellström, Å. (2012b). If you have a choice, you have trouble: Stimulus valence modulates presentation-order effects in preference judgment. *Journal of Behavioral Decision Making*, 25, 82–94. doi:10.1002/bdm.714
- Fechner, G. T. (1860). *Elemente der Psychophysik [Elements of psychophysics]*. Leipzig, Germany: Breitkopf & Härtel.
- Fechner, G. T. (1876). *Vorschule der Ästhetik [Introduction to aesthetics]*. Leipzig, Germany: Breitkopf & Härtel.
- Freimuth, M., & Wapner, S. (1979). The influence of lateral organization on the evaluation of paintings. *British Journal of Psychology*, 70, 211–218.
- Guilford, J. P. (1954). *Psychometric methods*. New York, NY: McGraw-Hill.
- Hellström, Å. (1977). Time errors are perceptual. *Psychological Research*, 39, 345–388.
- Hellström, Å. (1978). Factors producing and factors not producing time errors: An experiment with loudness comparisons. *Perception & Psychophysics*, 23, 433–444.
- Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 460–477.
- Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, 97, 35–61. doi:10.1037/0033-2909.97.1.35
- Hellström, Å. (1989). What happens when we compare two stimuli? In G. Ljunggren & S. Dornic (Eds.), *Psychophysics in action* (pp. 25–39). Berlin, Germany: Springer.
- Hellström, Å. (2000). Sensation weighting in comparison and discrimination of heaviness. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 6–17.
- Hellström, Å. (2003). Comparison is not just subtraction: Effects of time- and space-order on subjective stimulus difference. *Perception & Psychophysics*, 65, 1161–1177. doi:10.3758/BF03194842
- Houston, D. A., Sherman, S. J., & Baker, S. M. (1989). The influence of unique features and direction of comparison on preferences. *Journal of Experimental Social Psychology*, 25, 121–141. doi:10.1016/0022-1031(89)90008-5
- Jewell, G., & McCourt, M. E. (2000). Pseudoneglect: A review and meta-analysis of performance factors in line bisection tasks. *Neuropsychologia*, 38, 93–110. doi:10.1016/S0028-3932(99)00045-7
- Kellogg, W. N. (1931). The time of judgment in psychometric measures. *American Journal of Psychology*, 43, 65–86.
- Koh, S. D. (1967). Time-error in comparisons of preferences for musical excerpts. *American Journal of Psychology*, 80, 171–185.
- Koh, S. D., & Hedlund, C. W. (1969). Pair comparisons of musical excerpts: Preference bias of schizophrenics and normals. *Archives of General Psychiatry*, 21, 717–730.
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.
- Maeda, H. (1959). On the inhibitory effects of extrapolated stimuli in the successive comparison of brightness (I). *Japanese Journal of Psychology*, 30, 8–19.
- Masin, S. C., & Agostini, A. (1991). Attentional scanning and space errors. *Perception & Psychophysics*, 50, 285–289.
- Mattingley, J. B., Bradshaw, J. L., Nettleton, N. C., & Bradshaw, J. A. (1994). Can task specific perceptual bias be distinguished from unilateral neglect? *Neuropsychologia*, 32, 805–817. doi:10.1016/0028-3932(94)90019-1
- McCourt, M. E., Freeman, P., Tahmahkera-Stevens, C., & Chaussee, M. (2001). The influence of unimanual response on pseudoneglect magnitude. *Brain and Cognition*, 45, 52–63.

- McCourt, M. E., & Garlinghouse, M. (2000). Stimulus modulation of pseudoneglect: Influence of line geometry. *Neuropsychologia*, 38, 520–524.
- McLaughlin, J. P., & Kermisch, J. (1997). Salience of compositional cues and the order of presentation in the picture reversal effect. *Empirical Studies of the Arts*, 15, 21–27.
- Needham, J. G. (1935). The effect of the time interval upon the time-error at different intensive levels. *Journal of Experimental Psychology*, 18, 530–543. doi:[10.1037/h0056775](https://doi.org/10.1037/h0056775)
- Nicholls, M. E. R., Bradshaw, J. L., & Mattingley, J. B. (1999). Free-viewing perceptual asymmetries for the judgment of brightness, numerosity and size. *Neuropsychologia*, 37, 307–314.
- Patching, G. R., Englund, M. P., & Hellström, Å. (in press). Time- and space-order effects in timed discrimination of brightness and size of paired visual stimuli. *Journal of Experimental Psychology: Human Perception and Performance*. doi:[10.1037/a0027593](https://doi.org/10.1037/a0027593)
- Rhode, L., & Elias, L. J. (2007). Visual bisection of freely-viewed asymmetrical stimuli. *Laterality*, 12, 347–355.
- Rueckert, L., Deravanesian, A., Baboorian, D., Lacalamita, A., & Repplinger, M. (2002). Pseudoneglect and the cross-over effect. *Neuropsychologia*, 40, 162–173.
- Tant, M. L. M., Kuks, J. B. M., Kooijman, A. C., Cornelissen, F. W., & Brouwer, W. H. (2002). Grey scales uncover similar attentional effects in homonymous hemianopia and visual hemi-neglect. *Neuropsychologia*, 40, 1474–1481. doi:[10.1016/S0028-3932\(01\)00197-X](https://doi.org/10.1016/S0028-3932(01)00197-X)
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352. doi:[10.1037/0033-295X.84.4.327](https://doi.org/10.1037/0033-295X.84.4.327)
- Wänke, M. (1996). Comparative judgments as a function of the direction of comparison versus word order. *Public Opinion Quarterly*, 60, 400–409. doi:[10.1086/297761](https://doi.org/10.1086/297761)
- Wänke, M., Schwartz, N., & Noelle-Neumann, E. (1995). Asking comparative questions: The impact of the direction of comparison. *Public Opinion Quarterly*, 59, 347–372. doi:[10.1086/269481](https://doi.org/10.1086/269481)
- Woodrow, H. (1933). Weight discrimination with a varying standard. *American Journal of Psychology*, 45, 391–416.