EFFECTS OF TIME-ORDER ON PREFERENCE JUDGMENTS OF COLORS REPRESENTED BY PATCHES, LABELS, AND THEIR COMBINATION

Mats P. Englund
Department of Psychology, Stockholm University, SE-106 91 Stockholm, Sweden
mats.englund@psychology.su.se

Abstract

There are time-order error (TOE) studies for esthetic stimuli, but there seems to be an absence of studies comparing results from same versus different within-pair stimulus representations. In two experiments, participants made paired comparisons of colors represented by patches or labels (e.g., cherry red) (Exp. 1) and of their combination (Exp. 2), indicating within-pair preferences using a six-grade category scale. In Experiment 1, the two representations (patch-patch and label-label) were presented in two separate blocks, and in Experiment 2, the two within-pair orders (patch-label and label-patch) were randomized in one block. The results showed positive TOEs in pairs of patches, but not for labels, and there were positive TOEs in the combined color pairs. In terms of Hellström's sensation-weighting model, weights were greater for the second color than for the first in the comparisons of patches, labels, as well as combined stimuli.

Fechner (1860) first noted and named the phenomenon time-order error (TOE), the tendency to overestimate/underestimate the first of two successively presented stimuli. The TOE phenomenon is a curious one, and since Fechner's days, many theories have been suggested to explain it (for review, see, e.g., Hellström, 1985). One such theory is Hellström's (1979, 1985) *sensation weighting* (SW) model. This model has been shown to be successful in explaining TOEs for several different stimulus modalities, for example, tone loudness (Hellström, 1979), heaviness (Hellström, 2000), line length and auditory and visual duration (Hellström, 2003). According to the SW model, the comparison of two stimuli is not merely the subtraction of the stimulus magnitude of the second stimulus from that of the first. Instead, the subjective difference d is the difference between two weighted entities (subscripts 1 and 2 denoting first and second stimulus, respectively):

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

where k is a scale constant, s_1 and s_2 are weighting coefficients, ψ_1 and ψ_2 are the subjective stimulus magnitudes, $\psi_{\mathrm{r}1}$ and $\psi_{\mathrm{r}2}$ are magnitudes corresponding the current reference levels (ReLs; potentially different), and b is a constant accounting for effects not attributable to the weighting process (e.g., a judgment bias). The, often reasonable, assumption $\psi_{\mathrm{r}1} = \psi_{\mathrm{r}2} = \psi_{\mathrm{r}}$, coupled with $b = 0$ and $\psi_1 = \psi_2 = \psi$, reduces Equation 1 to

$$d = k(s_1 - s_2)(\psi - \psi_{\mathrm{r}}). \tag{2}$$

Equation 2 states that, for the weight relation $s_1 < s_2$, which is the most common, the overall TOE will be negative whenever the mean of the stimulus series is greater then the ReL.

So far, the vast majority of the TOE literature has focused on stimulus comparisons between stimuli varied on physical continua (e.g., weight or sound pressure level), and, seemingly, only a few studies have focused on comparisons of esthetic stimuli. Two such exceptions are the studies by Koh (1967) and Hellström (2001). In Koh's study, participants made paired comparisons of musical excerpts, choosing the more pleasant one. The results showed that TOE changed sign with the general pleasantness (as rated by other participants) of the excerpts in the respective pair. Specifically, for unpleasant excerpts there were positive TOEs, and for pleasant excerpts there were negative TOEs. Overall, there was a negative TOE, meaning that the second excerpt in a pair was considered more pleasant.

In two experiments, Hellström's (2001) participants made preference judgments, choosing the more pleasant stimulus in pairs of color patterns and pairs of jingles, respectively. In both experiments there were negative TOEs, which Hellström interpreted being due to the weight relation $s_1 < s_2$ and a stimulus series mean greater than the ReL (cf. Eq. 2). Hellström also suggested that this weight relation together with a ReL above the stimulus series explains the results of Koh's (1967).

Given the focus on actually perceived comparison stimuli in previous TOE studies, an important question is whether TOEs are limited to stimuli where the comparison depends on the actual perception of the stimuli at presentation, or if it also appears when the stimulus presentation only provides descriptions of stimuli to be imagined. The present study aimed to compare results of preference judgments in paired comparisons of similar stimuli, which were presented as either perceived stimuli, as labels, or their combination.

Experiment 1

Method

A total of 96 undergraduate students (21 men and 75 women) with a mean age of 26.8 (*SD* = 7.1) participated to fulfill a partial course requirement. The participant was seated in a dark room, 87 cm from a ViewSonic P227f monitor connected to a PC (Hewlett-Packard). Each preference trial consisted of a first stimulus (850 ms.), an ISI (1300 ms.), a second stimulus (850 ms.), and a request for a judgment represented by the question: *Which color do you prefer?* The experiment was sectioned into two preference blocks and two blocks for indicating general opinion on each stimulus. In the preference blocks, all colors were compared with all others, but themselves, in an order uniquely random for each participant. The general opinion ratings were always done after all preference judgments of the respective stimulus type; the presentation orders in the general opinion ratings were uniquely random for each participant. Within each block (i.e., color patches or color labels), the randomization had the restriction that two pairs constituted by the same two colors (i.e., A-B and B-A) were not presented within 20 pairs from each other. Preferences were indicated with a computer mouse on an on-screen six-grade category rating scale. The response alternatives were lined up vertically and centered on the screen with the following labels written on them: *Prefer the first color very much, Prefer the first color, Prefer the first color somewhat, Prefer the second color somewhat, Prefer the second color,* and *Prefer the second color very much*. For half of the participants, the response scale was reversed. The ratings of general opinion of the respective color were made on an on-screen seven-grade category rating scale with the response alternatives: *Very good, Good, Somewhat good, Neutral, Somewhat bad, Bad,* and *Very bad* lined up horizontally and centrally. The order of these alternatives was the same for every participant. Thus, a 2x2 factorial design was applied (scale order X stimulus-type order) (*Prefer first color very much* as topmost or bottommost alternative X patch-patch before label-label or vice versa), and each participant was randomly assigned to one of the four conditions. Before the experiment started, the participant was instructed that one step on the response scales was always an equally long step, regardless location and direction of the step. The whole experimental session took on average 32 minutes.

The color label letter height was 16.8 mm. The color patches (60 x 60 mm) were chosen to roughly match the color labels (CIE values *x*; *y*; luminance values in cd/m^2, in parentheses): *Cherry red* (.511; .300; 1.85), *Chocolate brown* (392; 375; 1.91), *Forest green* (x = 287; 537; 3.55), *Lemon yellow* (.416; .496; 90.3), *Lime green* (.345; .551; 78.3), *Mandarin orange* (.523; .409; 40.2), *Navy blue* (.161; .086; 1.27), *Salmon pink* (.495; .365; 38.7), *Sand Brown* (.400; .419; 13.9), *Sky blue* (.245; .323; 83.7), and *Tomato red* (.623; .328; 22.8). The background luminance of the screen was .35 cd/m^2.

Results and discussion

Equation 1 can be rewritten as

$$d = W_1\psi_1 - W_2\psi_2 + A, \tag{3}$$

where $W_1 = ks_1$, $W_2 = ks_2$, and $A = \psi_{r1} - \psi_{r2} + s_2\psi_{r2} - s_1\psi_{r1}$. Thus, the regression of the preference ratings on the general opinion ratings of the stimuli in the respective stimulus pair yields estimates of W_1 and W_2 as the regression coefficients. The regressions were highly significant in all cases, and the mean multiple *R* for color patches and color labels (henceforth, *patches* and *labels*) was .77 and .81, respectively.

The *W* values were entered into a repeated measures ANOVA (multivariate approach with Pillai tests) with stimulus position (first and second) and stimulus type (patches and labels) as within-subjects factors and stimulus-type order and scale-order as between-subjects factors. Two main effects and one interaction were significant: The weight for the second stimulus was greater than that of the first, $F(1, 92) = 27.31$, $p < .001$, the mean weight for the first and second stimulus was greater for labels than for patches, $F(1, 92) = 19.50$, $p < .001$, and the mean weight for the first and second stimulus was greater for labels than for patches for participants who compared patches before labels, $F(1, 92) = 16.09$, $p < .001$ (Figure 1).

Two repeated measures ANOVAs (multivariate approach with Pillai tests) were run for patches and labels, separately, with stimulus position as a within-subjects factor and stimulus-type order as a between-subjects factor. In both cases, there were only one significant effect, which was the main effect of stimulus position, $F(1, 94) = 19.59$, $p < .001$ and $F(1, 94) = 9.69$, $p < .01$, respectively.

Greater *W* values for labels than for patches suggest that it was easier to indicate preferences to the labels. Also, the interaction of stimulus-type order and stimulus type suggest that the participants who compared the labels first made use of the labels in the comparison of the patches. That is, those who compared labels first may very well have realized which patches were represented by which labels. They may then have attached those labels to the patches, thus making the comparisons easier for themselves by needing to rely less on their memory of the color actually presented on the screen. One important difference between comparing patches and comparing labels is that when comparing the patches, the participant needs to rely on the memory of what was actually presented on the screen. When comparing the labels, on the other hand, the participant can base the decision on the image of the color described by the label. In contrast to the memory of the patches, the memory of the label will not dissipate with time. Thus, by comparing the labels first, these participants had the benefit of being able to name the patches and thus make the comparison slightly easier.

TOE was calculated as the mean preference over all stimulus pairs. There were positive mean TOEs for both patches ($M = 0.015$) and labels ($M = 0.005$); although the TOEs

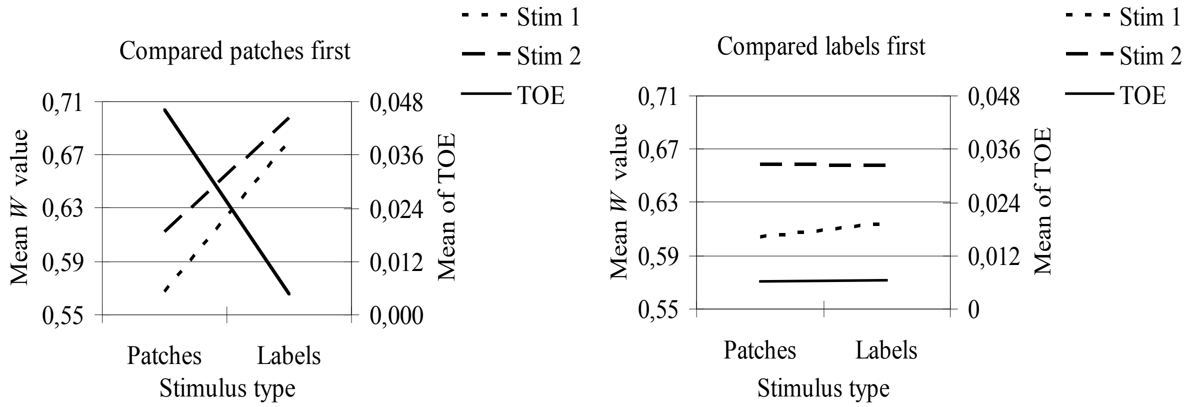

Figure 1. Mean *W* values (for first and second stimulus) and TOEs for patches and labels, separately for those who compared patches first and labels first, respectively.

were only significantly different from zero for patches, $t(95) = 2.08$, $p < .05$, and not for the labels, $t(95) = 1.01$, $p = .31$. The positive TOEs are contrary to what is usually found (e.g., Hellström, 1985, 2001; Koh, 1967). The mean general opinion rating was 0.42 for patches and 0.67 for labels; both statistics were significantly greater than zero, $t(95) = 8.86$, $p < .001$ and $t(95) = 13.75$, $p < .001$, respectively. With a positive TOE for patches and the weight relation $s_1 < s_2$, and assuming equal ReLs, Equation 2 implies that the ReL is greater than the stimulus series mean. Assuming equal ReLs and no bias, Equations 1 and 3 yield:

$$A = (W_2 - W_1)\psi_r. \qquad (4)$$

With data of sufficient quality, the mean ReL can be estimated as the slope in the regression of A on $(W_2 - W_1)$ through the origin. This regression yielded a mean ReL for patches of 0.26, $CI_{.95} = 0.01, 0.51$, multiple $R^2 = .043$. The meager R^2 value indicates the very low reliability of the estimation of the ReL. In fact, the confidence interval of the estimate of the ReL encompassed the entire confidence interval of the stimulus series mean ($CI_{.95} = 0.14, 0.49$). Therefore, no certain conclusion on the SW model's ability to explain the present data could be obtained. One explanation for the low reliability in the estimates is that, in contrast to classic TOE experiments, the measure of the stimulus magnitude is based on subjective ratings and is prone to variation between subjects. In addition, there was only one comparison per stimulus pair and, most importantly, only one general opinion rating per stimulus.

Experiment 2

Experiment 2 was designed to investigate what happens with the weighting of the stimuli when each pair is constituted by one stimulus of each type. Hence, in this experiment, every pair was constituted by one patch and one label, or vice versa.

Method

A total of 54 undergraduate students (10 men and 44 women) with a mean age of 27.4 (*SD* = 7.2) participated to fulfill a partial course requirement.

Apparatus, stimuli, and procedure were the same as in Experiment 1, with the following exceptions: Patches were compared with labels and labels with patches, all possible combinations of patch-label and label-patch were randomized uniquely in one single block for each participant, the response alternatives were in the same order (*Prefer the first color very* *much* as the topmost alternative) for all participants, and the colors were presented in one single block uniquely randomized for each participant for the ratings of general opinion.

Results and discussion

The regressions of the preferences on the general opinion ratings were highly significant in all cases and the mean multiple *R* was .72 for patch-label pairs and .72 for label-patch pairs.

The mean general opinion ratings for the patches (M=0.31) and labels (M = 0.95) were similar to those in Experiment 1. Both means were significantly greater than zero, $t(95) = 3.56$, $p < .001$ and $t(95) = 15.14$, $p < .001$, respectively, and significantly higher for labels than patches, $t(95) = 7.20$, $p < .001$.

The *W* values were entered into a repeated measures ANOVA (multivariate approach with Pillai tests) with stimulus position (first and second) and stimulus-type order (patches first and labels first) as within-subjects factors. The only significant effect was a main effect of stimulus order, $F(1, 53) = 17.93$, $p < .001$; the weight for the second stimulus was greater than that for the first (see Figure 2). TOE was calculated as the mean difference between the preferences in both presentation orders for each stimulus pair, across all stimulus pairs. There was a significantly positive TOE, $t(53) = 2.60$, $p < .02$ (see, Figure 2). The ReL for patch-label order was estimated (see Experiment 1) to 0.13, $CI_{.95}$ = -0.50, 0.75, multiple $R^2 = .003$, and for the label-patch order to 0.42, $CI_{.95}$ = -0.12, 0.96, multiple $R^2 = .043$. As in Experiment 1, it is not possible to test the SW model prediction that the stimulus series mean should be lower than the ReLs, due to the low reliability in the estimates of the ReLs.

General discussion

The main results presented here are the relation between the weights for the first and second stimulus in the paired comparisons. In both experiments, there was a greater weight for the second stimulus than for the first. This indicates that the participants took the second stimulus into more consideration than the first. Not only was this relation apparent for the both representations of the color stimuli (i.e., patch-patch and label-label), but also when stimulus pairs were constituted by both representations (i.e., patch-label and label-patch). Apparently, this relation is a persistent phenomenon found in most cases (e.g., Hellström, 1985, 2003). One explanation for the persistence of this phenomenon is that comparisons have a direction. That is, the comparison of A and B is either a comparison of A *to* B, or of B *to* A (e.g., Tversky, 1977). Tversky argued that in similarity judgment tasks defined as "How similar is

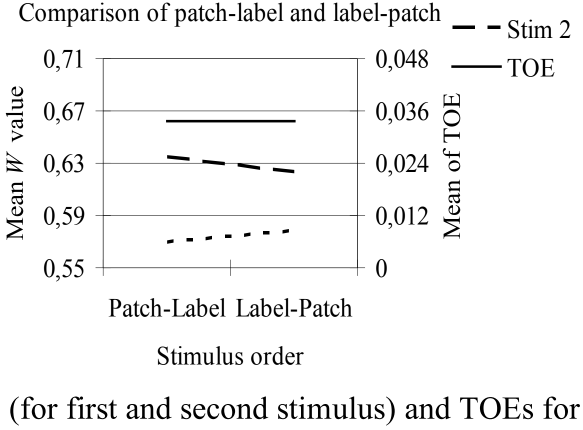

Figure 2. Mean *W* values (for first and second stimulus) and TOEs for the two stimulus orders patch-label and label-patch, respectively.

A to B?," the comparison has a direction so that A is subject and B is referent in the comparison. In many instances, however, the comparison direction is not prescribed in the instructions for the task. Instead, more direction neutral questions are often used, for example, *How similar are A and B?*, *Which tone was louder?*, or *Which do you prefer?* In these cases the comparison may very well be directed, not due to the instruction, but due to experimental conditions. One possible determinator of the direction of the comparison in a typical psychophysical experiment is the length of the ISI. For long ISIs, the weight relation is usually $s_1 < s_2$, but for short it is reversed (e.g., Hellström, 1979, 1985, 2003).

The proposition made here is that a reversal in the weight relation reflects a change in the direction of the comparison. Connecting this with Hellström's viewpoint, this means that the direction of the comparison is dependent upon the amount of information available for the respective stimulus at the point of decision.

Hellström (e.g., 1985, 1986, 2003) suggested that the purpose of the weighting in of the stimuli is to increase the signal-to-noise ratio so as to improve the discrimination between two stimuli when one of them changes. The weights for the respective stimulus reflect the amount of information the perceiver has of those stimuli at the point of making a judgment. The idea is that, at long ISIs, the memory trace of the first stimulus has dissipated more than that of the second stimulus. Thus will the second stimulus receive a greater weight than the first; the perceiver has to rely less on the reference level regarding the second stimulus. At short ISIs, on the other hand, the first stimulus is not completely processed before the second stimulus is presented. There is, therefore, partial interference between those stimuli, disturbing the processing of both stimuli. Because the processing of the first stimulus started before the interference, though, there is more information of the first stimulus available, and the weight of that stimulus is of a greater magnitude than that of the second stimulus.

It is, however, not possible to state from or to which direction the comparison changes. The weight for one stimulus being of greater magnitude than that for the other does not logically entail a specific comparison direction. This is an issue for future investigation.

References

Fechner, G.T. (1960). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig: Breitkopf & Härtel.

Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Perception and Performance, 5,* 460-477.

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

Hellström, Å. (1986). Sensation weighting in comparing: A tool for optimizing discrimination. In B. Berglund, U. Berglund, & R. Teghtsoonian (Eds.), Fechner Day 86 (pp. 89-94). Stockholm: International Society for Psychophysics.

Hellström, Å. (2000). Sensation weighting in comparison and discrimination of heaviness. *Journal of Experimental Psychology: Human Perception & Performance, 26,* 6-17.

Hellström, Å. (2001). Time-order effects for aesthetic preference. In G. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Fechner Day 2001. Proceedings of the Seventeenth Annual Meeting of the International Society for Psychophysics* (pp. 421-426). Lengerich, Germany: Pabst.

Hellström, Å. (2003). Comparison is not just subtraction: Effects of time- and space-order on subjective stimulus difference. *Perception & Psychophysics, 65,* 1161-1177.

Koh, S.D. (1967). Time-error in comparisons of preferences for musical excerpts. *The American journal of psychology, 80,* 171-185.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.