inappropriate for the required task, a time consuming translation is required. Arabic but not verbal numerals have privileged access to an analog-magnitude representation. A translation into a new representation is needed only with number words and not with Arabic numbers. Therefore performance is faster with Arabic numbers than with number words and the former are immune to interference from the latter in tasks involving numerical magnitude.

References

Arieh, Y., & Algom, D. (2002). Processing picture-word stimuli: The contingent nature of picture and of word superiority. Journal of Experimental Psychology: Learning, *Memory and Cognition, 28*, 221-232.

Damian, M. F. (2004). Asymmetries in the processing of Arabic digits and number words. *Memory & Cognition, 32*, 164-171.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1-42.

Fias, W., Reynvoet, B.,& Brysbaert, M. (2001). Are Arabic numerals processed as pictures in a Stroop interference task? *Psychological research, 4*, 242-249.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

SENSATION WEIGHTING IN PREFERENCE JUDGMENTS CREATES A GOODNESS-LEVEL DEPENDENT WORD-ORDER EFFECT

Mats P. Englund

Department of Psychology, Stockholm University, SE-106 91 Stockholm, Sweden

mats.englund@psychology.su.se

Abstract

Hellström (2003) found a goodness-level dependent word-order effect (WOE) for preference judgments. However, Hellström used preference scales adapted to each stimulus pair, and goodness was rated for whole stimulus pairs. Therefore, participants might not have judged preference, and statistical testing of the stimulus weights was not possible. In the present study, 211 participants indicated within-pair preferences for 25 stimulus pairs by choosing one of six written (same for all stimuli) preference expressions (e.g., "apple I like more than *pear"). Within-pair presentation order was reversed for half of the participants. Participants also rated each stimulus's goodness by choosing one of seven written expressions (e.g., "Apple* I generally like*"). Results replicated Hellström's results; there was a greater weight for the left stimulus and a positive correlation between WOE and goodness level. These results strengthen the evidence that the WOE is due to differential sensation weighting rather than, for instance, semantic congruity.*

Two physically equivalent stimuli separated by time or by space are often reported as subjectively different on some appropriate continuum. Fechner (1860) named these respective phenomena time-order error (TOE) and space-order error (SOE), defined as positive (negative) when the first/left (second/right) stimulus is overestimated relative to the other. Since Fechner's days the literature on TOEs and SOEs has grown greatly, and now spans order effects on a wide range of psychophysical continua. Many factors have been found to influence the sign and magnitude of TOEs and SOEs (see, e.g., Hellström, 1985 for review). For example, TOEs in esthetic comparisons varied linearly with pleasantness (Koh, 1967), and SOEs in comparisons of line lengths changed from negative to positive when stimulus magnitude varied from low to high at longer durations (Hellström, 2003a).

Plenty of analogous examples of choice-alternatives' presentation orders' affecting the outcome of judgments have been reported in cognitive psychology. For example, presentation order influenced answers to poll-type questions (Wänke, Schwarz, & Noelle-Neumann, 1995) as well as consumer-type choices between alternatives presented as written descriptions (Houston & Sherman, 1995). However, researchers in cognitive psychology do not seem to have noted the psychophysical literature on order effects. In particular, the dependence of the magnitude and sign of TOEs and SOEs on stimulus magnitude does not seem to have been acknowledged in the cognitive psychology literature. In fact, Hellström (2003b) seems to have been the first to investigate the possibility that the sign and magnitude of order effects depend on stimulus levels in choices between stimuli represented by verbal descriptors (i.e., everyday objects and phenomena).

In Hellström's (2003b) study, participants made preference judgments by agreeing to one of six written preference statements for each of 10 stimulus pairs and rated their general opinion on the stimuli. The results showed a positive linear relation between a *word-order* *effect* (WOE; a positive WOE is an overestimation of the left/first read stimulus) and the general goodness level of the stimulus pairs. That is, there was a tendency to prefer the leftmost/first read (henceforth left) alternative of two good stimuli and the rightmost of two negative stimuli. Hellström explained these effects with his *sensation weighting* (SW) model (e.g., Hellström, 1979), which describes the subjective difference between two stimuli as

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

where k is a scale constant, s_1 and s_2 are weighting coefficients, ψ_1 and ψ_2 are subjective stimulus magnitudes, ψ_{r1} and ψ_{r2} are magnitudes corresponding current reference levels (ReLs; potentially different). The WOE equals d when $\psi_1 = \psi_2$; assuming $\psi_{r1} = \psi_{r2} = \psi_r$ Equation 1 reduces to

$$\text{WOE} = d = k(s_1 - s_2)(\psi - \psi_r). \quad (2)$$

Thus, according to Equation 2, the dependence of the WOE on the goodness level in Hellström's (2003b) study was caused by the weight relation $s_1 > s_2$ (subscripts 1 and 2 denoting left and right stimulus).

A few points may be made regarding Hellström's (2003b) study. First, the preference scales were adapted to fit the specific stimulus pairs (e.g., apple is much *tastier* than pear; and childcare is more *important* than eldercare), which suggests that participants perhaps were not making *preference* judgments. Second, the goodness level ratings were made for whole pairs instead of individual stimuli. The weight difference Hellström used to explain his results was thus inferred from the slope of the WOE regressed on goodness level. Therefore, if the apparent goodness-level dependent WOE was an artifact of, for example, semantic congruity (Banks, Clark, & Lucy, 1975, a tendency to choose "better" from two good stimuli and "worse" from two bad stimuli), the weight difference could also have been an artifact. Though TOEs are not caused by the semantic congruity effect (Hellström, 1985; Petrusic & Baranski, 1989), Hellström statistically tested semantic congruity as a potential cause of the observed effect and concluded that a semantic congruity explanation was implausible. Even so, data confirming the weight relation ($s_1 > s_2$) by direct statistical testing would be desirable as further evidence against such a semantic congruity explanation for the goodness-level dependent WOE. A third point is, as Hellström discussed, that several participants misunderstood the goodness-rating task and that some of the stimulus selections may have been inadequate. The purpose of the present study was to conceptually replicate the Hellström study, but using a single preference scale, collecting ratings on general goodness level for individual stimuli, attempting to construct an easier-to-understand goodness-rating task, and using a primarily different stimulus selection with a larger stimulus set.

Method

A total of 211 undergraduate psychology students (46 men and 165 women) with a mean age of 26.4 (*SD* = 6.2) participated to fulfill a partial course requirement. Participants received a three-section booklet; i) a preference judgment task, ii) a filler task (two personality tests) to conceal the true purpose of the study, and iii) a stimulus goodness-rating task. In the preference section, there were three stimulus pairs to be compared on each page. For each stimulus pair there were six preference statements written on individual lines, each starting and ending with the stimuli to be compared (A "…I like [*much more*, *more*, *somewhat more*, *somewhat less*, *less*, *much less*] than…" B). Centered directly under each preference statement was a short dashed line; the participants were to make their preference judgments by marking

with an "X" the dashed line under the preference statement that agreed the most with their own opinion. Half of the participants received a reversed within-pair word order (i.e., B-A instead of A-B). In the stimulus goodness rating task, the setup was similar to that of the preference judgment task. For each stimulus there were seven statements ("… I generally [*like greatly*, *like*, *like somewhat*, *neither like nor dislike*, *dislike somewhat*, *dislike*, *dislike greatly*] …"). To the right of each goodness statement, there was a short dashed line aligned with the right margin. Participants were to make their goodness ratings by marking with an "X" the dashed line to the right of the goodness statement they agreed the most with.

Results

Due to incompletely filled out booklets, data from 12 participants were removed before analyses were made. Preference ratings were scaled as "2.5" (strongest preference for the left) to "-2.5" in steps of one. The goodness ratings were analogously scaled from "3" (highest positive goodness) in steps of one to "-3." Means and standard deviations for preference ratings and goodness ratings for each stimulus pair and the two respective word orders are presented in Table 1.

According to Equation 1, the subjective difference on the preference continuum can, for the two presentation orders, be written as (Hellström, 2003b)

$$d_{AB} = k\{[s_1\psi_A + (1 - s_1)\psi_{r1}] - [s_2\psi_B + (1 - s_2)\psi_{r2}]\}; \quad (3)$$

$$d_{BA} = k\{[s_1\psi_B + (1 - s_1)\psi_{r1}] - [s_2\psi_A + (1 - s_2)\psi_{r2}]\}. \quad (4)$$

Using group data, the WOE can thus be formulated as

$$\text{WOE} = (d_{AB} + d_{BA}) / 2. \quad (5)$$

According to Equation 5, taking half the sum of the relative preferences for the two groups (AB and BA) for each stimulus pair gives the within-pair WOEs. The mean WOE over all stimuli ($M = 0.026$) was not statistically significant, $t(23) = 0.95$; $p = .35$. However, when the WOE values were plotted against the mean goodness levels of the stimulus pairs, a fitted regression line revealed a statistically significant positive linear relationship between the WOE and the goodness level (Figure 1). Equation 1 can be rewritten as

$$d = W_1\psi_1 - W_2\psi_2 + A, \quad (6)$$

where $W_1 = ks_1$, $W_2 = ks_2$, and $A = k(\psi_{r1} - \psi_{r2} + s_2\psi_{r2} - s_1\psi_{r1})$. The regression of the preference judgments on the goodness ratings of the respective stimulus in each pair yields estimates of W_1 and W_2 as the regression coefficients. The WOE values were regressed on the group means of goodness for the left and right stimuli, respectively ($R = .858$; $W_1 = 0.510$; and $W_2 = 0.435$). The difference in magnitude of W_1 and W_2 was tested with a regression with the WOE as dependent variable and goodness difference and sum, respectively, of left and right stimulus as predictors. The regression coefficient of the goodness difference was significant, $t(22) = 4.58$, $p < .0001$, indicating that W_1 was significantly greater than W_2. The goodness-level dependent WOE (Figure 1) can thus be explained by $s_1 > s_2$ and Equation 2.

As seen in Figure 1, stimulus pair 11 (Licorice-Chocolate) deviated mysteriously from the other pairs and the regression line. Inspection of Table 1 shows that the difference in mean goodness level between the groups was extraordinarily large (0.59) for this pair. A closer look at the data revealed that the mean difference between the two stimuli was different

Table 1. *Means (*SD*s in parentheses) of preference and goodness ratings for stimulus pairs.*

Stimulus pair *i* (A_i-B_i)	Preference (A_i-B_i) Group AB	Preference (A_i-B_i) Group BA	Goodness of pairs Group AB	Goodness of pairs Group BA
1. Aftonbladet-Expressen	0.66 (1.07)	-0.75 (1.08)	-0.12 (1.54)	-0.41 (1.54)
2. Cinema-Concert	-0.21 (1.47)	0.64 (1.58)	2.13 (1.02)	2.11 (0.96)
3. Blood pudding-Veal brawn	0.98 (1.59)	-0.99 (1.62)	-1.01 (1.79)	-0.96 (1.79)
4. Finland-Norway	-0.56 (1.36)	0.61 (1.12)	0.93 (1.08)	0.95 (1.13)
5. Wasps-Mosquitoes	-0.27 (1.72)	-0.18 (1.63)	-1.90 (1.14)	-2.05 (1.03)
6. Cough-Runny nose	-0.59 (1.44)	0.32 (1.22)	-2.09 (0.95)	-2.13 (0.84)
7. Dogs-Cats	0.28 (1.32)	-0.02 (1.65)	1.86 (1.30)	1.50 (1.58)
8. Headache-Stomachache	0.20 (1.34)	-0.47 (1.35)	-2.51 (0.72)	-2.54 (0.71)
9. High jumping-Long jumping	0.38 (1.32)	0.07 (1.32)	0.63 (1.18)	0.34 (1.04)
10. Fall-Winter	0.51 (1.60)	-0.39 (1.63)	0.87 (1.48)	0.86 (1.51)
11. Licorice-Chocolate	-1.39 (1.43)	1.03 (1.42)	1.62 (1.46)	2.07 (1.18)
12. Deciduous forest-Coniferous forest	0.90 (1.15)	-0.65 (1.22)	1.76 (1.00)	1.69 (1.08)
13. Snakes-Spiders	0.13 (1.80)	-0.14 (1.59)	-1.46 (1.48)	-1.58 (1.38)
14. Pasta-Rice	0.67 (1.40)	-0.26 (1.28)	1.78 (1.48)	1.82 (0.95)
15. Piano-Guitar	0.22 (1.43)	0.07 (1.38)	1.90 (1.02)	1.82 (1.17)
16. Plastic carrier-Paper carrier	-0.35 (1.36)	0.34 (1.41)	0.63 (1.10)	0.64 (1.09)
17. Pripps-Carlsberg	-0.88 (1.16)	0.75 (1.18)	0.43 (1.42)	0.47 (1.36)
18. Forest green-Sea blue	-0.61 (1.34)	0.92 (1.46)	1.78 (0.99)	1.86 (1.02)
19. Filth brown-Dust gray	-0.57 (1.26)	0.51 (1.22)	-0.49 (1.25)	-0.33 (1.30)
20. Swedish Television-TV4	0.22 (1.60)	-0.26 (1.52)	1.16 (1.25)	1.19 (1.17)
21. Toothache-Nausea	0.00 (1.41)	-0.29 (1.34)	-2.68 (0.57)	-2.61 (0.75)
22. News on TV-News in paper	-0.13 (1.20)	0.12 (1.20)	1.48 (0.96)	1.55 (0.92)
23. Bus-Subway	0.82 (1.52)	-0.46 (1.40)	0.71 (1.31)	0.80 (1.20)
24. Apple-Pear	0.58 (1.24)	-0.28 (1.37)	1.82 (0.90)	1.71 (1.07)
Mean and *SD* of means	0.04 (0.61)	0.01 (0.54)	0.39 (1.55)	0.37 (1.56)

Note. Stimulus pairs 1, 17, and 20 denote two major Swedish evening newspapers, two beer brewing companies, and two major Swedish TV channels, respectively.

in groups AB and BA (-1.86 and -0.79, respectively). For group AB, the mean goodness of Licorice and Chocolate were 0.69 and 2.55, respectively, and for group BA 1.67 and 2.46. It seems, therefore, that the deviation of Licorice-Chocolate is explained by a coincidentally less favorable opinion on licorice in group AB.

To test the weighting effects at the individual level, preference ratings were regressed on the goodness of each stimulus in the respective pair for each participant. The mean *R*, over participants, was .73 ($R^2 = .55$). The *W* values were entered into a repeated-measures ANOVA (multivariate approach) with stimulus position (first and second) as a within-subjects factor and word order (AB vs. BA) as a between-subjects factor. The main effect of stimulus position was significant, $F(1, 197) = 53.40$, $p < .001$, indicating that $s_1 > s_2$ ($W_1 = 0.75$, $W_2 = 0.67$, a weight difference of 11%). The interaction term was not significant, $F(1, 197) = 0$.

Combining Equations 5 and 6 yields the effect of goodness level on the WOE as

$$\text{WOE} = [(W_{AB1} + W_{BA1})\psi_1 - (W_{AB2} + W_{BA2})\psi_2 + (A_{AB} + A_{BA})] / 2, \quad (7)$$

thus WOE = $0.746\psi_1 - 0.668\psi_1 - 0.008$. The WOEs for “greatly disliked,” “neither liked nor

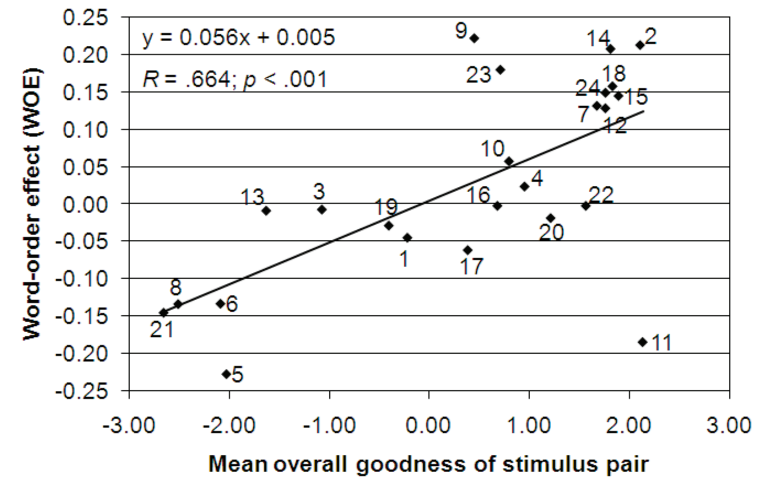

Figure 1. WOE plotted against mean goodness level of stimulus pairs (labels from Table 1). A fitted regression line and its equation are shown. A positive WOE means a tendency to prefer the left stimulus. (Omitting pair 11 changes *R* to .812 and the equation to $y = 0.066x + 0.016$).

disliked,” and “greatly liked” stimuli were -0.24, 0.01, and 0.23 scale steps, respectively. These values multiplied by 2 give the preference difference between the presentation orders.

Discussion

The main results presented here are the goodness-level dependent WOE and the stimulus weight relation $s_1 > s_2$. These results replicate those of Hellström (2003b), and demonstrate the robustness of these effects. The present results strengthen the evidence for a sensation weighting explanation for the goodness-level dependent WOE, rather than, for example, semantic congruity (Banks et al., 1975), in particular because it was shown explicitly that the weight of the left stimulus was of greater magnitude than that of the right. Furthermore, the present results validate the effects on a preference continuum because all preference judgments were made on a single preference scale. Here, most stimuli were different from those used by Hellström, which generalizes the WOE further. The practical importance of the WOE seems real; for a stimulus pair at the extreme ends, the preference difference is on average approximately half a scale step.

The deviation of the stimulus pair Licorice-Chocolate (11) seemed mysterious at first glance, but seems to be explained by the coincidentally lower opinion of licorice in group AB. The effect of presentation order for this pair was likely concealed by an unequal advantage of chocolate over licorice for group AB as compared to group BA. Exclusion of pair 11 increased *R* to .81, which is equal to what Hellström (2003b) found.

One might ask if the WOE should be treated as a TOE, a SOE, or as its own effect. It should be noted that Englund (2007) found TOEs in preference comparisons of color stimuli, but only when at least one of the stimuli in the pairs was represented by a color patch. No TOEs were found when both stimuli were represented by labels (e.g., *lime green*). The results of Englund further suggest that the TOE is dependent upon memory processes (cf., e.g., Hellström, 1985). In a typical SOE experiment, two stimuli are presented briefly and simultaneously and separated in space, and the judge has to rely greatly on the memory of the two stimuli just presented. In the present study, however, both stimuli were present throughout the entire decision process, even though they were read in a particular order. This puts doubt on any straightforward comparison of WOEs to TOEs and SOEs. Therefore, it seems that the term WOE is still the most appropriate.

In cognitive psychology, a widespread theoretical explanation of preference reversals builds on Tversky's (1977) contrast model of similarity (e.g., Houston & Sherman, 1995). This explanation assumes that a choice involves a directional comparison of a *subject* to a *referent*. The comparison is an active search for the subject's features in the referent, where shared features are cancelled and unique features of the subject decide the choice. Because the comparison is assumed to be directional, this means that when the subject has unique positive features, it will be chosen due to the active search for the subject's features in the referent. When the subject has unique negative features, the referent will be chosen due to the salience of the subject's negative features. The evidence for feature matching has mainly come from studies where stimuli were presented as structured feature lists (e.g., Houston & Sherman, 1995), and the stimuli shared positive but not negative features or vice versa. A feature-matching model seems limited to judgments where alternatives can be separated easily into sets of features and inappropriate in, for example, esthetic preference judgment tasks. The SW model is a more general comparison model that does not suffer from dependence on stimulus presentation formats. The present results do not, however, rule out feature matching. A natural direction for future studies is to test these two models against one another.

References

Banks, W. P., Clark, H. H., & Lucy, P. (1975). The locus of the semantic congruity effect in comparative judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *104*, 35-47.

Englund, M. P. (2007). Effects of time-order on preference judgments of colors represented by patches, labels, and their combination. In S. Mori, T. Miyaoka, & W. Wong (Eds.), *Fechner Day 07* (pp. 251-256). Tokyo, Japan: International Society for Psychophysics.

Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig: Breitkopf & Härtel.

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 errors and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, *97*, 35-61.

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

Hellström, Å. (2003b). Word-order effects in preference comparison. In B. Berglund & E. Borg (Eds.), *Fechner Day 2003. Proceedings of the 19th Annual Meeting of The International Society for Psychophysics, Larnaca Bay, Cyprus, October 19-23, 2003.* Stockholm: International Society for Psychophysics (pp. 121-126).

Houston, D. A., & Sherman, S. J. (1995). Cancellation and focus: The role of shared and unique features in the choice process. *Journal of Experimental Social Psychology*, *31*, 357-378.

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

Petrusic, W. M., & Baranski, J. V. (1989). Semantic congruity effects in perceptual comparisons. *Perception & Psychophysics*, *45*, 439-452.

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

Wänke, M., Schwartz, N., & Noelle-Neumann, E. (1995). Asking comparative questions: The impact of the direction of comparison. *Public Opinion Quartely*, *59*, 347-372.

TIME- AND SPACE-ORDER EFFECTS IN STIMULUS COMPARISON IN THE LIGHT OF RESPONSE-TIME DATA

Åke Hellström

Department of Psychology, Stockholm University, SE-106 91 Stockholm, Sweden

hellst@psychology.su.se

Abstract

As has been found experimentally, magnitude comparison of paired successive or simultaneous stimuli can be described as being based on differential subtractive weighting of the stimulus-induced sensations ψ_1 *and* ψ_2*, causing time- and space-order effects (TOEs and SOEs). In the simplest case, the comparison equation becomes*: $d = s_1 (\psi_1 - \psi_r) - s_2 (\psi_2 - \psi_r)$, *where* ψ_r *is the reference level (ReL). New analyses of response-time data from stimulus comparison experiments (Hellström, 2003) show that* signed response speed*, the inverse of response time with the sign of the subjective difference d, carries similar information on* d *as measures based on transformed proportions of "1st greater," "equal," and "2nd greater" responses. Sensation weighting is similar for fast and slow responses, which suggests that it arises in a preceding processing stage, and strengthens the notion that the weighting, and thereby TOEs and SOEs, is of perceptual origin.*

Hellström (2003) reported five experiments where paired successive or simultaneous stimuli were compared in order to study time- and space-order effects. In these experiments response times (RTs) were recorded but not utilized. Here, the use of RTs in response scaling and analyses of results is discussed. Results from three of the experiments are used for illustration.

Experimental Method

Each participant took part in one session with 1–5 self-paced computerized experiments in irregular order, indicating the longer, louder, etc. stimulus by pressing, with the right hand on a numeric keyboard, "1" (1st or left), "2" (2nd or right), "0" (cannot decide), and then "Enter."

Each experiment had four 24-pair sets, each with its separate ISI or stimulus duration but the same stimulus combinations. All 96 were presented in a pseudorandom order, the same for all participants. Data from three of the experiments are used here.

Experiment 3 (Line length, successive; 39 participants). Stimuli were horizontal orange lines, 2 mm thick, centered on blue background, 100 ms in duration. The first line (Base) had six equal steps within 96 mm ± 60%, and the second line four equal steps within Base ± 22%. ISIs were 100, 300, 900, and 2,700 ms.

Experiment 4 (Tone loudness, successive; 41 participants). Stimuli were 600-Hz tones, 100 ms in duration. The Base tone had four equal sone steps within 11 sones [75 dB(A)] ± 60%, and the second tone six equal steps within Base ± 20%. ISIs were as in Exp. 3.

Experiment 5 (Line length, simultaneous; 31 participants). Stimuli were horizontal orange lines, 2 mm thick, on blue background. A centered orange fixation cross appeared 3 s before the stimuli, lasting 2 s. The lines were centered 96 mm to each side. Their durations were 100, 200, 400, or 800 ms. Equally often the left and the right line was the Base, with three equal steps within 48 mm ±60%. The other line had four equal steps within Base ±13%.