Journal of Experimental Psychology:
Human Perception and Performance
1982, Vol. 8, No. 4, 582-601

0096-1523/82/0804-0582$00.75

Loci of Contextual Effects in Judgment

Barbara A. Mellers
University of California, Berkeley

Michael H. Birnbaum
University of Illinois at Urbana-Champaign

Three experiments investigated the loci of contextual effects in judgment. Experiment 1 demonstrated the effect of stimulus spacing on category ratings and magnitude estimations of the darkness of dot patterns. Variations in the stimulus spacing were shown to affect both category ratings and magnitude estimations in a similar fashion. Experiment 2 was designed to determine whether contextual effects due to stimulus spacing influence the scale values or the judgment function. Subjects judged "differences" and "ratios" of the subjective darkness of dot patterns. Differences in mean judgments of single stimuli from Experiment 1 did not predict the rank order of judged "differences" and "ratios" from Experiment 2. The estimated scale values of the stimuli appeared to be independent of stimulus spacing. These findings suggest that contextual effects due to the stimulus spacing occur in the judgment function for within-modality judgments. Experiment 3 examined contextual effects in cross-modality judgments. Stimulus spacing and stimulus range were manipulated for "difference" and "total" judgments. Unlike the within-modality results, the stimulus range and spacing influenced the scale values. A contextual theory of within- and cross-modality judgment is presented.

It is now well established that judgments are relative. That is, the response to a stimulus depends not only on the stimulus to be judged, but also on other stimuli that form a *context*, or frame of reference for judgment (Birnbaum, 1974b, 1978; Helson, 1964; Johnson & Mullally, 1969; Parducci, 1963, 1968, 1974, 1982; Poulton, 1968; Restle & Greeno, 1970). The term *context* refers to such factors as the stimulus range, stimulus spacing, frequency of stimulus presentation, response range, and other experimental details that influence judgment.

What is not so well established is how to deal with contextual effects (Anderson, 1975; Birnbaum, 1974b, 1978; Poulton, 1979). Some authors have regarded contextual effects as sources of error, noise, confusion, or bias. Stevens (1971), for example, remarked that they "rate no better than a nuisance" and are a "diversion from the basic business of sorting out the fundamental principles" (p. 448). Poulton (1979) and Anderson (1982) contended that contextual effects can and should be "avoided," though their recommendations for procedures that are supposed to accomplish this goal differ.

Others have argued that contextual effects are both lawful and necessary for a complete psychophysical theory. This approach attempts to develop theories that account for judgments in a wide variety of contexts (Birnbaum, 1974b, 1982a; Parducci, 1974, 1982; Birnbaum, Parducci, & Gifford, 1971). The purpose of this paper is to explore contextual effects in within-modality and cross-modality comparison and combination tasks.

Outline of Judgment

To facilitate discussion of possible loci of contextual effects, it is useful to represent judgment as a composition of functions. In single-stimulus judgments, such as category ratings or magnitude estimations, subjective values (s) of the stimuli are related to physical values by the psychophysical function, $s = H(\phi)$. The output function, J, translates subjective values to overt responses. The function relating responses to stimuli is the

---

Coauthorship is equal.

This research was supported by the Research Board of the University of Illinois at Urbana-Champaign. We thank Jerome R. Busemeyer for discussion of an earlier draft.

Requests for reprints should be sent to either Barbara A. Mellers, Department of Psychology, University of California, Berkeley, California 94720 or Michael H. Birnbaum, Department of Psychology, University of Illinois, 603 East Daniel, Champaign, Illinois 61820.

composition, $R = J[H(\phi)]$. Judgments based on multiple stimuli (judgments of "differences," "ratios," or "total intensities") can be represented as the composition of three functions (Birnbaum, 1974a, 1978), as in Figure 1.[1]

Contextual effects due to manipulation of the stimuli that form the frame of reference can, in principle, influence any or all of these functions. Identifying the loci of contextual effects amounts to pinpointing contextual effects in terms of these functions. A complete psychophysical theory should not only identify the loci of contextual effects but successfully predict changes in the function (or functions) influenced by manipulation of the context.

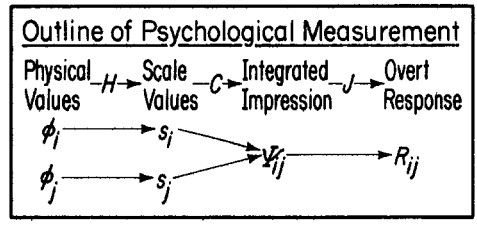

Figure 1. Outline of notation. (Subjective scale values of the stimuli, s_i and s_j are combined [or compared] by the function, $\psi_{ij} = C[s_i, s_j]$, and transformed to an overt response by the strictly monotonic judgment function, $R_{ij} = J[\psi_{ij}]$. From Birnbaum, 1978.)

Range-Frequency Theory

One contextual theory for predicting category ratings of single stimuli as a function of the stimulus distribution is range-frequency theory (Parducci, 1965; 1982; Parducci & Perrett, 1971). This theory assumes that the response is a compromise between two tendencies: (a) a tendency to divide the stimulus range into equal subranges; and (b) a tendency to use categories along the response scale with equal frequency. A general form of range-frequency theory can be written:

$$G_{ik} = \alpha_k F_k(s_i) + \beta_k s_i + \gamma_k , \qquad (1)$$

where G_{ik} is the response to stimulus i in context k; s_i is the subjective value of the stimulus; F_k is the cumulative stimulus distribution for context k (i.e., $F_k[s_i]$ is linearly related to the rank of stimulus i in context k); α_k and β_k are the weights that depend on the response range, stimulus range, and weights of the two tendencies, and γ_k is an additive constant that depends on the response scale. If α_k is zero, then the response will be linearly related to the subjective value of the stimulus (range principle). On the other hand, if β_k is zero, then the response will be linearly related to the stimulus rank (frequency principle). Range-frequency theory has been successful in a wide variety of judgment domains (Birnbaum, 1974b; Parducci, 1963; 1974; Parducci & Perrett, 1971). Range-frequency theory may apply to contextual effects in the psychophysical function or the judgment function (H or J in Figure 1).

Overview

The purpose of this article is to identify the loci of contextual effects in within-modal and cross-modal judgments. All three experiments use the same stimuli: dot patterns. The stimulus levels (in number of dots) were spaced according to either a positively or negatively skewed distribution in which six dot patterns were common to all distributions and to all three experiments. Judgments or inferred scale values of these common stimuli are affected by the other stimuli in their distribution; this influence is termed a *contextual* effect.

In Experiment 1, judges rated or estimated the "darkness" of single-dot patterns using category ratings or magnitude estimations. In Experiment 2, judges evaluated "ratios" and "differences" in darkness between each pair of stimuli. In Experiment 3, judges made cross-modality comparisons and combinations of dot patterns with circles that varied in diameter. They were asked to judge the "difference" between the darkness of each dot pattern and the size of each circle and the "total intensity" of each pair.

[1] Quotation marks are used to designate the task given to the subject or responses given by the subject. Quotation marks are not used for models or theoretical statements. Thus "ratios" and "differences" refer to judgments of stimulus pairs, which may or may not fit ratio and difference models. Quotations are not used for computed differences in judgment, which are actual differences between two judgments.

Experiment 1 investigates the effects of the stimulus spacing and the response procedure. In Experiment 1, contextual effects can be attributed to the composition of two functions, $J[H(\phi)]$. With single stimulus judgments, it is impossible to unconfound contextual changes in the scale values (H) from changes in the judgment function (J).

Experiment 2 separates the effects of the stimulus spacing on H and J by examining two-stimulus "ratio" and "difference" judgments. Previous research (Birnbaum, 1978, 1980) concluded that the subtractive operation underlies both "ratio" and "difference" judgments. An extension of the subtractive theory which allows variations in the stimulus spacing to affect *both* the scale values and the judgment function, can be written:

$$R_{ijk} = J^*_k(s_{jk} - s_{ik}), \tag{2}$$

$$D_{ijk} = J_k(s_{jk} - s_{ik}), \tag{3}$$

where R_{ijk} and D_{ijk} represent "ratio" and "difference" judgments, respectively; J^*_k and J_k are judgment functions that depend on the response procedure and the distribution of stimuli, k; s_{ik} and s_{jk} are the estimated scale values for stimulus i and j that also depend on the stimulus distribution, where $[s_{ik} = H_k(\phi_i)]$. Experiment 2 investigates Equations 2 and 3 as well as two special cases, in which contextual effects influence only the scale values *or* only the judgment functions. These special cases make different predictions that will be tested in Experiment 2.

Experiment 3 examines the effects of the stimulus range and spacing on cross-modal judgments. Subjects are asked to compare and combine the subjective darkness of dot patterns with the subjective size of circles. A general theory that asserts that contextual effects both precede (occur in H) and follow (i.e., in J) stimulus comparison can be expressed by the equations

$$T_{ijk} = J'_k(c_j + s_{ik}), \tag{4}$$

$$D_{ijk} = J_k(c_j - s_{ik}), \tag{5}$$

where T_{ijk} and D_{ijk} are the "total intensity" and "difference" judgments, respectively; J'_k and J_k are strictly monotonic judgment functions; s_{ik} and c_j are the scale values of the ith dot pattern and the jth circle size in context k. (Because the same distribution of circle sizes is used throughout, the circle scale values do not have a subscript for context.) Special cases of Equations 4 and 5 that assume that the stimulus range and spacing operate entirely on H or J make differential predictions that will be tested in Experiment 3.

The major purpose of these experiments is to test alternative theories of the loci of contextual effects in which the influence of such variables as stimulus spacing can be attributed to different functions in Figure 1. It will be shown that the effects of context can be attributed to the J functions in Experiments 1 and 2. However, in the cross-modality tasks of Experiment 3 there was evidence that scale values depend on stimulus spacing and range.

Experiment 1: Contextual Effects in "Direct" Scaling

Traditionally, the two most popular "direct" methods for scaling have been category ratings and magnitude estimations. Magnitude estimations are usually found to be a positively accelerated function of category ratings (Stevens, 1966; Stevens & Galanter, 1957; Torgerson, 1961; Eisler, 1963; Marks, 1974, 1979). The nonlinear relationship between the two procedures has been a long-standing puzzle.

One theory is that judgments of stimuli can be regarded as "direct" measurements of subjective value. This theory ignores the J function in Figure 1 and operationally defines scale values as judgments:

$$G_{ik} = s_{ik}, \tag{6}$$

$$M_{ik} = s^*_{ik}, \tag{7}$$

where G_{ik} and M_{ik} are category ratings and magnitude estimations of stimulus i in context k, s_{ik} and s^*_{ik} are the subjective values, in which a nonlinear transformation, $s_{ik} = T(s^*_{ik})$, relates the two scales.

An alternative representation of these judgments (Birnbaum, 1978; 1982a) is as follows:

$$G_{ik} = J_k(s_i), \tag{8}$$

$$M_{ik} = J^*_k(s_i), \tag{9}$$

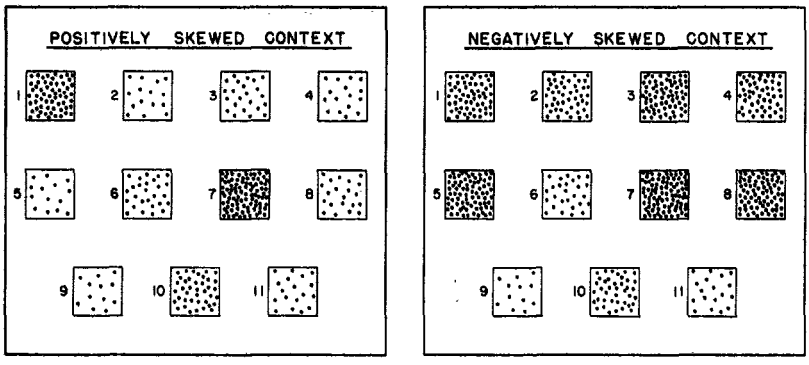

Figure 2. Two stimulus distributions. (Patterns 9, 11, 6, 10, 1, and 7 are identical in both contexts; these stimuli have 12, 18, 27, 40, 60, and 90 dots, respectively.)

where s_l is assumed to be a sensory scale value, independent of the stimulus range, stimulus spacing, or the response scale; J_k and J_k^* are assumed to be strictly monotonic judgment functions whose exact functional form depends lawfully on the stimulus range, stimulus spacing, and response continuum.

There are three major purposes for Experiment 1. First, the experiment determines the magnitude of the contextual effect due to stimulus spacing using the same stimuli and general procedure that will be used in Experiment 2. Predictions for Experiment 2 will be calculated based on results of Experiment 1. The second purpose of Experiment 1 is to examine the effects of stimulus spacing and response range for both category ratings and magnitude estimations. The same stimuli and procedures are used to determine whether the effects are comparable for the two tasks. A third purpose of Experiment 1 is to investigate the consequences of comparing magnitude estimations between groups of subjects who experience different contexts. Equations 8 and 9 imply that these judgments are not necessarily an ordinal scale of subjective value when compared across contexts.

Method

Observers were asked to judge the subjective darkness of dot patterns using either category ratings or magnitude estimations. Four anchored category-rating conditions were produced by a factorial design of two response ranges combined with two stimulus distributions. In two other category-rating conditions, the largest and smallest stimuli were not anchored to the endpoints of the response scale, to examine the effect of anchoring. Twelve magnitude-estimation conditions were produced by a factorial combination of two stimulus distributions, three standards, and two response ranges. Different subjects served in each of the 18 conditions, which were carried out over a two-year period as separate experiments with the same stimuli.

Stimuli. The stimuli were 25-mm squares containing irregular patterns of 1-mm solid dots. The judges received either a positively skewed or a negatively skewed distribution of dot patterns, as shown in the left and right of Figure 2, respectively. Six stimuli, common to both distributions, contained 12, 18, 27, 40, 60, and 90 dots (patterns numbered 9, 11, 6, 10, 1, and 7 in Figure 2, respectively). In the positively skewed distribution, five additional (context) stimuli had 14, 15, 16, 21, and 23 dots. In the negatively skewed distribution the five context stimuli had 47, 51, 70, 74, and 77 dots.

Each subject received a 22 cm × 28 cm sheet on which the 11 stimuli were printed in the format shown on either side of Figure 2, except that there were no titles or border, and each stimulus was identified by a letter instead of a number. Thus, all of the stimuli for one context were simultaneously available during the judgments.[2] The common stimuli appear in the same positions in both contexts so that position in the arrangement and stimulus rank would be unconfounded.

Category rating conditions. The positively and negatively skewed stimulus distributions were factorially combined with three response scales. In two of the conditions, judges were asked to use integers from 1 to 5 where 1 = very light, 2 = light, 3 = medium, 4 = dark, and 5 = very dark. In two other conditions, the rating scale went from 1 to 100 (where 1 = very light and 100 = very dark). Rating scales were anchored by the instructions to call the lightest dot pattern 1 and the darkest dot pattern either 5 or 100. In two other con-

[2] Parducci (1963) found that contextual effects are similar for situations in which the stimuli are presented successively or simultaneously. Birnbaum (1978, 1982a) found similar results for the two procedures for judgments of "ratios" and "differences." In pilot work, we obtained similar results to our Experiment 2 with successive presentation of stimulus pairs.

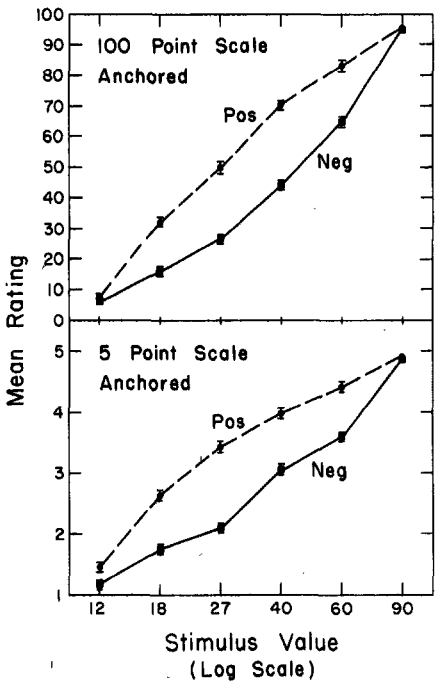

Figure 3. Contextual effects in category ratings. (Mean ratings of the common stimuli are plotted against stimulus values, which are spaced in equal log steps. Dashed curves show ratings for the positively skewed stimulus distribution [left of Figure 2]; solid curves are from the negatively skewed distribution [right of Figure 2]. Upper panel shows results for 1–100 rating scale; lower panel for 1–5 rating scale. Brackets show plus and minus one standard error.)

ditions (positively and negatively skewed stimulus distributions), the 1–5 scale was used, but the category labels were not anchored to the stimuli.

The subjects repeated the experiment in all of the conditions, and only the second set of responses was used in the analyses.

Magnitude estimation conditions. The stimuli, distributions, and procedures were the same as those for the category-rating tasks except that judges were asked to make magnitude estimations. Positively and negatively skewed stimulus distributions were factorially combined with two ranges of response examples and three standards (12, 18, and 27 dots), yielding 12 conditions.

Judges were asked to estimate the "ratio" of each stimulus to the standard, using a modulus of 100. One set of instructions used the following examples: 33 = ⅓ as dark as the standard dot pattern, 40 = ⅖ as dark, 50 = ½ as dark, 67 = ⅔ as dark, 100 = equal in darkness to the standard, 150 = 1½ times as dark, 200 = 2 times as dark, 250 = 2½ times as dark, and 300 = 3 times as dark.

In the other instructions, the examples read: 11 = 1/9 as dark as the standard, 14 = 1/7 as dark, 20 = 1/5 as dark, 33 = ⅓ as dark, 100 = equal in darkness to the standard, 300 = 3 times as dark, 500 = 5 times as dark, 700 = 7 times as dark, and 900 = 9 times as dark. When the 12-dot pattern (lightest stimulus) was used as the standard, the examples below 100 were omitted.

In all of the magnitude estimation conditions, judges were encouraged to use any numbers, in between or more extreme than the examples, to indicate the "ratio of the darkness of each dot pattern to the standard."

Subjects. The judges were 883 undergraduates.[3]

In the category-rating conditions, there were 134 judges who used the 5-point anchored scale, 228 with unanchored 5-point scale, and 158 with the 100-point anchored scale. About half of each group received the positively or negatively skewed distributions.

In the magnitude estimation conditions, there were 363 subjects with 22 to 41 judges in each of the 12 conditions.

Results and Discussion

Category Ratings. Mean judgments for the anchored category-rating tasks are shown in Figure 3. The upper panel shows mean ratings of the common stimuli in the anchored 1–100 condition; the lower panel shows the results for the anchored 1–5 condition. The two curves within each panel show that the mean ratings of the same stimuli can be either positively or negatively accelerated relative to log ϕ, depending on the stimulus spacing. Brackets represent plus and minus one standard error for each mean. The general shape of the curves is consistent with Parducci's range-frequency theory (Equation 1). Data for the 1–5 unanchored scale were similar to those for the anchored scale.

Parducci (1982) has recently found that the magnitude of the contextual effect due to the stimulus distribution decreases with increasing number of response categories and that it increases as a function of the number of distinct stimulus levels. Using a 100-point scale, Parducci (1982) found a small contextual effect due to variation of the relative frequency with which 5 stimuli

[3] The research participants in Experiments 1–3 were 1,123 undergraduates at the University of Illinois at Urbana-Champaign, who were tested alone or in small groups and who received credit in lower division psychology courses. An additional 51 were tested who either failed to follow instructions or to complete the tasks in time. Of these, 33 students (most of whom failed to complete one of the two parts in the allotted time) were in Experiment 3; 13 were in Experiment 1; and the other 5 were in Experiment 2.

were presented. However, the present data, obtained with 11 stimulus values, show that the contextual effect of stimulus spacing for the 100-point rating scale remains quite large.

Magnitude estimations. Magnitude estimations of the common stimuli are shown in Figure 4, with a separate curve for each of the four conditions in which a standard of 12 dots was used. If there were no effect of the stimulus distribution and the examples used in the instructions, then all four curves would coincide. Instead, the difference between the open and solid points shows that when the examples range as high as 900, subjects use numbers that average much higher than when the largest example is only 300. It appears that the power function exponent obtained in magnitude estimation experiments depends largely on two variables that are under the experimenter's control: the (log) response range and the (log) stimulus range. Thus, exponents of the power function may relate more closely to the experimental design than to the subjects' sensations. Robinson (1976) and Poulton (1979) reached similar conclusions (see also Teghtsoonian, 1971).

Figure 4 shows that the effects of the stimulus spacing on magnitude estimations are similar to the effects for category ratings. Parducci (1963) found similar results. One difference is that the magnitude estimation curves in Figure 4 do not rejoin at the upper end, as they do for category ratings.[4]

Since the relationship between M and ϕ and between G and ϕ can have so many different forms, the relationship between M and G should not be theorized to be an invariant functional form (Montgomery, 1975).

In Figure 5, mean magnitude estimations of the same six stimuli (averaged over context) are shown for the between-subject designs (on the left) from Experiment 1 and a within-subject design from Experiment 2 (on the right; these data will be discussed in more detail later). In the between-subject designs, each subject is given only one standard and several comparison stimuli; in the within-subject design, each subject receives multiple standards and stimuli. The stimuli are spaced on the abscissa according to judgments using the 27-dot standard, which then automatically becomes the identity line in each set. The ratio model ($R_{ij} = s_j/t_i$, where R_{ij} is the magnitude estimation of the "ratio" of stimulus levels j to i, with scale values s_j and t_i) implies that the other curves in Figure

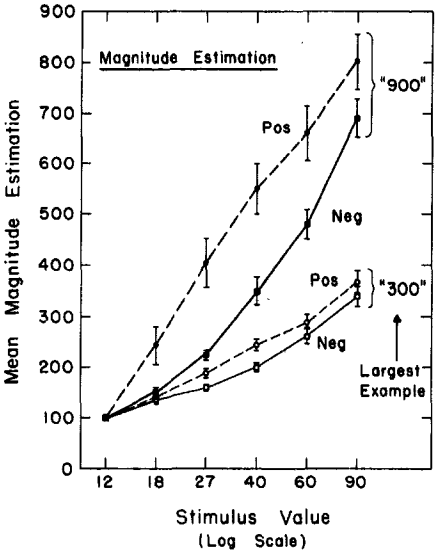

Figure 4. Contextual effects in magnitude estimation. (Mean magnitude estimations for common stimuli are plotted against stimulus values as in Figure 3. The 12-dot stimulus was the standard and was assigned the value 100. Dashed lines show results for positively skewed distributions, solid lines are for negatively skewed conditions. Upper two curves show results when instructions included an example response as high as 900; lower two curves show results when the highest example was 300. Brackets show plus and minus one standard error.)

[4] The present data allow a test of the conclusion of Moskowitz (1982) that magnitude estimations are more sensitive to stimulus differences than category ratings. Following the procedure of Moskowitz, the F ratio for the main effect of the common stimuli (relative to the Subjects × Stimulus interaction) was calculated for the first 22 subjects in each of the 18 conditions. Contrary to Moskowitz, it was found that the Fs for category ratings are larger than those for magnitude estimations. All 6 category-rating conditions had $F(5, 105)$ greater than 80, and 5 of 6 Fs were greater than 120. For the 12 magnitude-estimation conditions, only 6 of 12 had $F(5, 105)$ greater than 80, and only 2 Fs were greater than 120. A difference of procedure may have produced the difference in results. Moskowitz did not anchor his category scale to the stimuli by means of warm-ups and/or instructions to judge the stimuli relative to the extremes. It seems likely that his subjects assigned several stimuli to the same category and effectively used only a few categories.

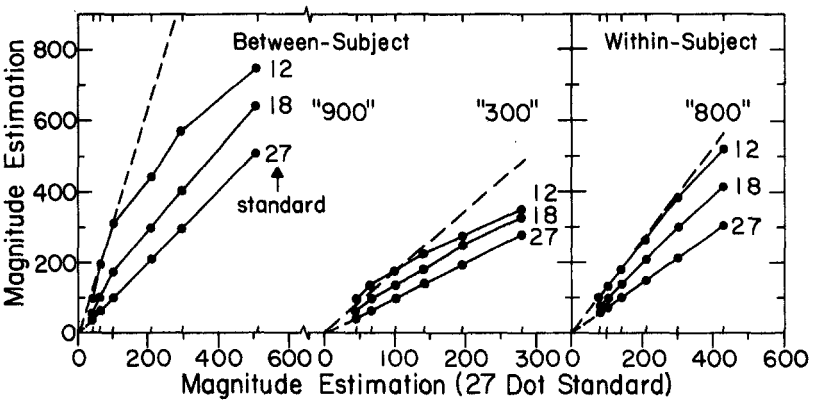

Figure 5. A comparison of magnitude estimations from between-subject designs (in which each judge received only one standard) and a within-subject design, plotted as a function of judgments using the 27-dot standard, with a separate curve for each standard. (Examples in the instructions went as high as 900 for the condition on the left, 300 for the curves in the center, and 800 for the condition on the right. Judgments of common stimuli were averaged over the positively and negatively skewed distributions. Dashed lines show predictions from the product rule for the upper curve in each set.)

5 should also be linear, and all three curves in each set should intersect at a common point. The two sets of curves obtained using between-subject designs (left of Figure 5) bow out in the center and pinch in at the ends. The curves obtained using a within-subject design show bilinear divergence, more consistent with the ratio model.

Product rule. Previous investigations of magnitude estimations have examined the product rule, which is implied by a special case of the ratio model.[5] The product rule can be written as

$$R_{ij} = R_{ik} \cdot R_{kj}, \qquad (10)$$

which follows from the ratio model, $R_{ij} = s_j/t_i$, if $s_j = t_i$ when $i = j$ (that is, if the scale value of first and second stimuli are equal for equal stimuli).

The dashed lines in Figure 5 show the prediction of the product rule (Equation 10). Only a portion of the dashed line is plotted for the curves on the left because the predicted curve goes beyond the scale on the ordinate. In fact, the predicted "ratio" of 90 dots to 12 dots is 1,593, although the mean response is only 746. The between-subject data clearly violate the product rule. The upper curve with the 12-dot standard is not as steep as the curve predicted by the product rule for either the "300" or the "900" condition. However, the within-subject data (Experiment 2) appear reasonably consistent with the product rule. The violations of the product rule and violations of bilinearity in the between-subject data are expected from Equations 8 and 9, which assume that different J and J^* functions are involved for different groups of subjects receiving different stimulus distributions, standards, and examples.

In sum, both category ratings and magnitude estimations of the subjective darkness of dot patterns depend on the stimulus spacing. Both procedures also depend on the response range presented in the instructions. In the case of magnitude estimations, it appears that the supposed freedom of the subject to control the range of responses is largely illusory, since subjects seem to be extremely sensitive to the range of examples used to explain the task.

[5] Bilinearity is a weaker requirement, implied by a more general ratio model, $R_{ij} = (s_j/t_i)^m + b$. If the curves intersect with an ordinate projection of zero, then $b = 0$. Otherwise, the ordinate projection of the intersection (b) can be subtracted from each "ratio" response. If $s_j = t_i$ when $i = j$, then these adjusted "ratio" responses ($R_{ij} - b$) should obey the product rule. A recent article by Fagot (1979) distinguishes three properties of ratio models, which he defines as *ratio consistency*, $R_{ij} = R_{ik} \cdot R_{kj}$; *product constancy*; $R_{ij} \cdot R_{jk} = R_{im} \cdot R_{mk}$; and *ratio constancy*, $R_{ij}/R_{kj} = R_{im}/R_{km}$. The first property is Equation 10. Violation of bilinearity would refute all three properties.

Experiment 2: Contextual Effects in "Ratio" and "Difference" Judgments

One interpretation of the results from Experiment 1 is that the judgment function depends upon the stimulus spacing and other contextual aspects of the experiment, $G_{ik} = J_k(s_i)$, where s_i is independent of context. However, another interpretation is that these results reveal changes in the psychophysical function, as expressed by the equation, $G_{ik} = s_{ik} = H_k(\phi_i)$. Experiment 2 is designed to separate the H function from the J function and thereby identify the loci of contextual effects due to stimulus spacing.

In Experiment 1, the difference in the responses between 18 and 12 dots exceeds the difference in the responses between 90 and 60 dots for the positively skewed stimulus distribution. However, for the negatively skewed distribution, the mean difference in the responses to 18 versus 12 is *less* than the mean difference in responses to 90 versus 60 (see Figures 3 and 4). With a unifactor design (as in Experiment 1) it is impossible to tell whether these changes in response should be attributed to changes in the scale values or changes in the judgment function.

In Experiment 2, however, it *is* possible to distinguish between contextual effects that precede or follow stimulus comparison. The experimental procedure involves asking subjects to *judge* "differences" between stimulus pairs and varying the spacing of the stimuli. For example, will judges in the positively skewed context rank the "difference" in darkness between 18 and 12 dots as greater than the "difference" in darkness between 90 and 60? Will the subjects given the negatively skewed distribution rank order the two "differences" in the opposite order? (Similar questions can be asked concerning "ratios.")

This question can be formalized with respect to two special cases of Equations 2 and 3. First, contextual effects may operate entirely *prior* to stimulus comparison; that is, they operate on the H function in Figure 1. This theory assumes that judgments of single stimuli directly reflect scale values, and s_{ik} can be replaced with judgments from Experiment 1, G_{ik}; that is, $s_{ik} = G_{ik}$. This theory implies that judged "differences" in Experiment 2 will be monotonically related to computed differences in judgment from Experiment 1 as follows:

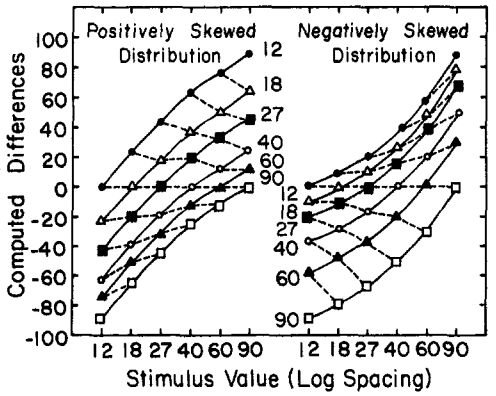

Figure 6. Computed differences between mean category ratings (on the 100-point scale) from Figure 3. (If stimulus spacing affects scale values, the rank orders of the judged "differences" in the two contexts would not be the same. Dashed lines, which connect equal physical ratios, go in opposite directions in the two panels.)

$$R_{ijk} = J^*[G_{jk} - G_{ik}], \qquad (2a)$$

$$D_{ijk} = J[G_{jk} - G_{ik}], \qquad (3a)$$

where R and D are "ratio" and "difference" judgments; J and J^* are monotonic judgment functions, as in Equations 2 and 3; G_{ik} represents the judged value of stimulus i in context k, as in Experiment 1.

Second, the effect of context may only *follow* stimulus comparison; that is, context may only operate on the J function in Figure 1. This second special case can be written as follows:

$$R_{ijk} = J_k^*[s_j - s_i], \qquad (2b)$$

$$D_{ijk} = J_k[s_j - s_i], \qquad (3b)$$

where only the judgment functions (J) depend on the context (subscript k). This theory implies that the rank order of "differences" and "ratios" will not vary as a function of stimulus spacing.

Predictions of the model in which context effects precede stimulus comparison (Equations 2a and 3a), are shown in Figure 6. Differences ($G_{jk} - G_{ik}$) were computed from the single stimulus category ratings in Experiment 1 (on the 1–100 anchored scale). These computed differences are plotted as

a function of the left stimulus (minuend—along the abscissa) with a separate curve for each level of the stimulus on the right (subtrahend). The left and right of Figure 6 show predicted differences computed from category ratings in the positively skewed and negatively skewed conditions, respectively.

Notice that the rank order differs systematically between the two conditions in Figure 6. Dashed lines connect pairs of equal physical ratios (i.e., equal log differences) to highlight the change in rank order. In the positively skewed condition, equal log differences *decrease* in subjective extremity as one moves up the scale. In the negatively skewed condition, equal log differences are predicted to *increase* in extremity as one moves up the scale. For example, Figure 6 shows that in the positively skewed context, the computed difference between the 18- and 12-dot stimuli *exceeds* the difference between the 90- and 60-dot stimuli, whereas for the negatively skewed context, the order of these differences is reversed.

If contextual effects govern H or both H and J, then the rank order of judged "differences" in the positively skewed context would be different from the rank order of the judged "differences" in the negatively skewed context. If the pattern in Figure 6 is obtained, the simpler interpretation that scale values are independent of context (Equations 2b and 3b) would be rejected in favor of Equations 2a and 3a or Equations 2 and 3. On the other hand, if the scale values are *independent* of the stimulus spacing—that is, if contextual effects operate only on the transformation from subjective differences to overt responses—then Equations 2b and 3b (in which s_j and s_i replace s_{jk} and s_{ik}) would be sufficient. This model implies that the rank order of "difference" and "ratio" judgments in both contexts should be the same: All four matrices should have the same rank order.

Method

Observers were asked to judge either "ratios" or "differences" of the subjective darkness of dot patterns. Positively and negatively skewed stimulus distributions (shown in Figure 2) were factorially combined with instructions to judge either "ratios" or "differences." Different judges served in each of the four conditions.

"Difference" task. Judges were told to rate the "difference of subjective darkness" between stimulus pairs on a scale from 90 to −90 where 80 = left stimulus is very very much darker than right; 60 = left stimulus is very much darker than right; 40 = left stimulus is much darker than right; 20 = left stimulus is darker than right; 0 = left and right stimuli are equal in darkness; −20 = right stimulus is darker than left; −40 = right stimulus is much darker than left; −60 = right stimulus is very much darker than left; −80 = right stimulus is very very much darker than left. Judges were instructed to use integers between −90 and 90.

"Ratio" task. Judges were given the following examples: 800 = left stimulus is 8 times as dark as right; 400 = left stimulus is 4 times as dark as right; 200 = left stimulus is 2 times as dark as right; 100 = left and right stimulus are equal in darkness; 50 = left stimulus is ½ as dark as right; 25 = left stimulus is ¼ as dark as the right; 12.5 = left stimulus is ⅛ as dark as right. Judges were encouraged to use numbers in between or more extreme than the examples to express their judgments.

Design. In each condition there were 121 trials constructed from an 11 × 11, Left Stimulus × Right Stimulus, factorial design in which each dot pattern on the left could appear with each of the 11 dot patterns on the right. The distribution of stimuli was the same for both the right and left stimulus.

Procedure. Judges read the instructions and completed 33 representative warm-up trials followed by 121 experimental trials in random order. Each trial referred to two of the dot patterns of the stimulus array. The stimulus sheets were the same as those used in Experiment 1. Thus, the stimuli and procedure were like those of Experiment 1 except that judges were directed to compare stimuli in pairs rather than judge them individually.

Subjects. There were 80 judges, with 18 to 23 different people in each of the four conditions.

Results and Discussion

Figure 7 shows mean "ratios" plotted against mean "differences" with a separate type of symbol for each divisor (subtrahend). Data are shown for the 6 × 6 common design, with the results for the positively skewed condition on the left, and the results for the negatively skewed condition on the right. Note that the data appear reasonably consistent with the hypothesis that one operation underlies both tasks, that is, that judged "ratios" are approximately a monotonic function of judged "differences." If subjects truly used both ratio and subtractive operations with the same scale values, however, then judgments of "ratios" and "differences" would *not* be monotonically related but instead would show a particular nonmonotonicity in which equal ratios cor-

respond to more extreme differences as the divisor increases. (see Birnbaum, 1980, Figure 3).

To test the one-operation theory and to find out if scale values depend on context, the judgments were fit by a special case of Equations 2 and 3 in which J and J^* are linear and exponential, respectively:

$$\hat{R}_{ijk} = \gamma_k \exp[s_{jk} - s_{ik}] + \delta_k , \quad (11)$$

$$\hat{D}_{ijk} = \alpha_k[s_{jk} - s_{ik}] + \beta_k , \quad (12)$$

where $\hat{R}_{ijk}$ and $\hat{D}_{ijk}$ are predicted "ratio" and "difference" judgments between stimuli i and j in context k; α_k, β_k, γ_k, and δ_k are fitted constants. The subscript, k, on the scale values indicates that different scale values were permitted for each context (though the scale values and comparison process were assumed to be the same for both "ratio" and "difference" tasks within each contextual distribution). Although Equations 11 and 12 assume that the scale value of each stimulus is independent of the standard with which it is compared and also that the scales are the same for both tasks, these equations gave a good fit to data from nine experiments with several continua (Birnbaum, 1978, 1980, 1982a).

For each task, a proportion of variance unaccounted for was defined as follows:

$$P_T = \frac{\Sigma\Sigma(X_T - \hat{X}_T)^2}{\Sigma\Sigma(X_T - \bar{X}_T)^2} , \quad (13)$$

where P_T is the proportion of residual systematic variance for task T, X_T is the mean judgment within each cell of the design, $\hat{X}_T$ the corresponding prediction, and $\bar{X}_T$ the overall mean judgment for task T. In the "difference" tasks, X_T represents the mean judgment; for the "ratio" tasks, X_T is the log of the mean judgment, $\hat{X}_T$ is the log of the prediction, and $\bar{X}_T$ is the mean of the logs. The summation is over all cells in the design for task T. The sum of these four proportions (across all four matrices) was minimized, by means of a computer program that utilized Chandler's (1969) STEPIT subroutine. Parameter estimates were derived using the 11×11 designs. Similar results were obtained when only the 6×6 common designs were used. For the common design, the overall indices of lack of fit were .011 and .014 for the positively and negatively skewed conditions, respectively. Model deviations thus constitute a little more than .5% of the systematic variance for each of the four matrices.

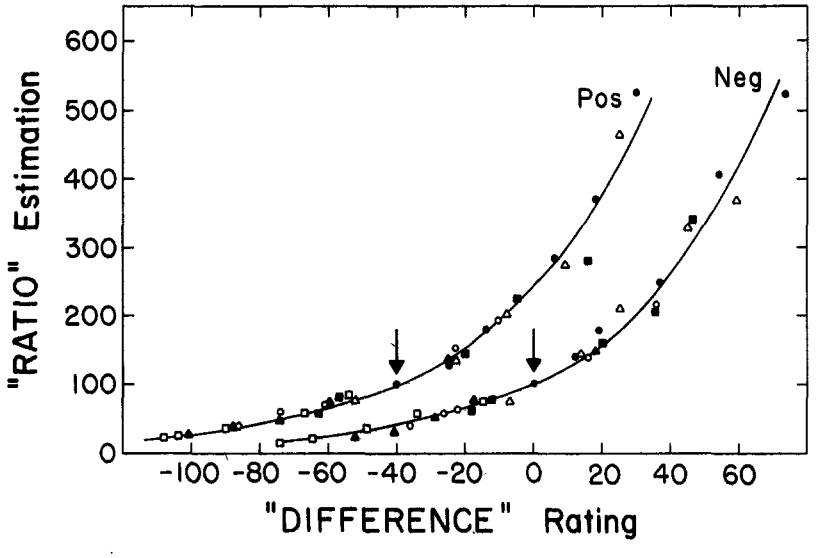

Figure 7. Judgments of "ratios" plotted against "differences." (Data on the left and right are for the positively and negatively skewed contexts, respectively. Abscissa shows scale for negatively skewed context; positively skewed context data are shifted 40 units to the left; arrows show zero "difference." Curves are best-fit solutions to a special case of the one operation theory.)

The fit of the model can be assessed in Figure 8, which shows mean judgments for the common stimuli in the positively skewed distribution (upper panels) and negatively skewed distribution (lower panels). Data points are connected with solid lines. Dashed lines are predictions of the theory that subjects use the same operation and scale values for both tasks (Equations 11 and 12).

The subtractive model (Equations 11 and 12) predicts that the "difference" judgments will be parallel and linear and that the "ratio" judgments will show bilinear divergence due to the exponential J function. In all four panels, the data points lie close to the predictions of the model. The average standard errors of the means are 2.55 and 8.62 in the "difference" and "ratio" tasks, respectively.

Note that not only are the rank orders of the data points similar within a context (i.e., for "difference" and "ratio" judgments) but also across contexts (i.e., for positive and negative skew). Thus, the predictions in Figure 6, which were based on judgments of the same stimuli in the same contexts (Experiment 1), failed to materialize. Instead, the

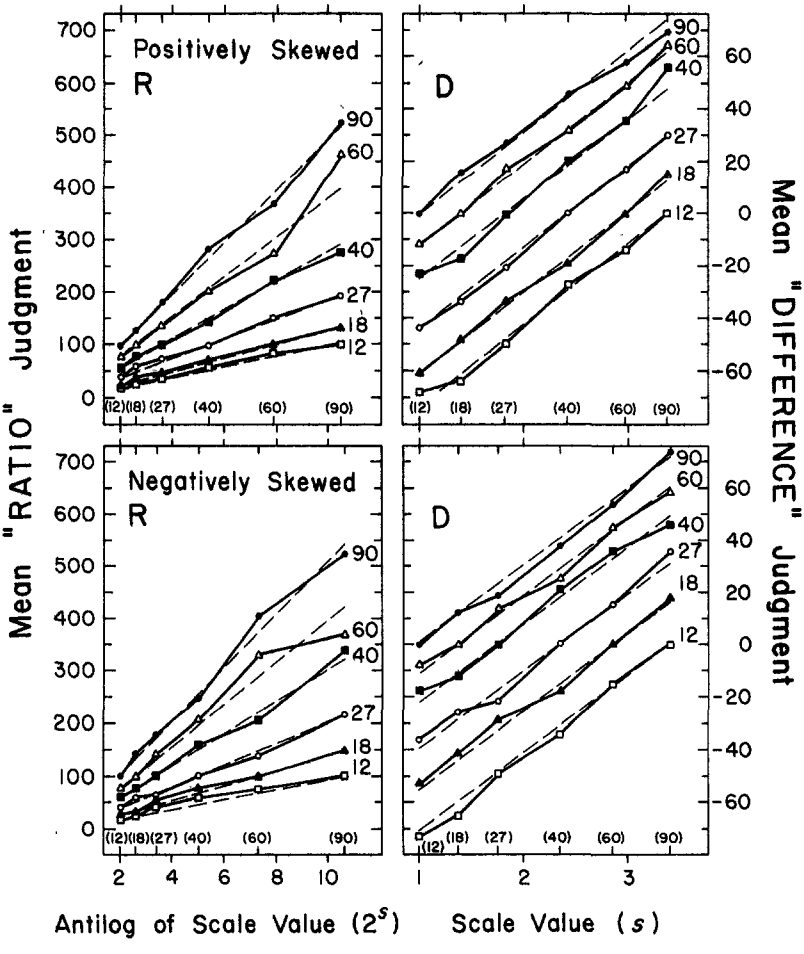

Figure 8. Mean "ratio" judgments (R) and "difference" judgments (D) for the common stimuli in the positively skewed and negatively skewed conditions. (This figure replots the data of Figure 7 to assess the predictions of bilinearity and parallelism. Dashed lines are the predictions of the subtractive theory fit to all four matrices simultaneously. "Ratios" are plotted against antilogs of the scale values for the left stimulus, with a separate curve for each stimulus on the right [divisor]. Linearity of the curves is consistent with the theory that one operation underlies both tasks.)

rank order of "differences" and "ratios" appears largely independent of stimulus spacing.

Figure 9 shows the estimated scale values for the positively skewed distribution plotted against those for the negatively skewed context. Note that although the scale values of the two conditions were permitted to differ, they are quite similar; the points lie very close to the identity line. The dashed line shows the expected relationship if the stimulus spacing had influenced the estimated scale values in the same fashion as it affected the single ratings of Experiment 1, according to Equations 2a and 3a (i.e., contextual effects precede stimulus comparison). Instead, the data appear consistent with the view that contextual effects operate only on the *J* function, as in Equations 2b and 3b.

In conclusion, these data do not provide evidence that contextual effects due to stimulus spacing operate on the scale values. It appears that the rank order of "ratios" and "differences" can be reproduced by assuming that judges compute differences between scale values that are independent of the stimulus spacing. Considering Experiments 1 and 2 together, it appears that judgments of single stimuli depend on stimulus spacing, but scale values derived from the subtractive model do not. Thus, it seems reasonable to localize the effects of stimulus spacing in the judgment function for "direct" ratings and for the present within-modal stimulus comparisons.

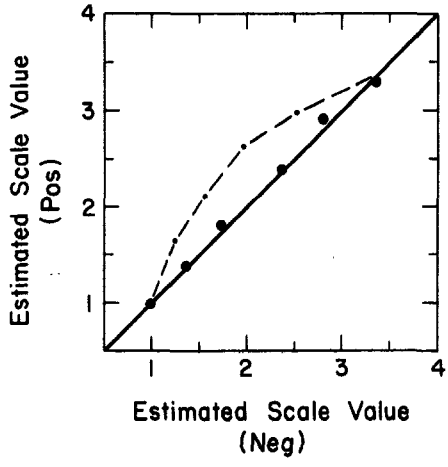

Figure 9. Solid points show estimated scale values for positively skewed context plotted against estimated scale values for negatively skewed context. (Broken curve shows relationship predicted from scale values based on the 100-point ratings in Figure 3.)

Experiment 3: Contextual Effects in Cross-Modality Comparison and Combination

Experiment 3 examines whether the same theories can account for contextual effects in cross-modality judgments as for within-modality judgments. Cross-modality matching involves the comparison of stimuli from two different dimensions. "Does the punishment fit the crime?" or "Is this salary fair pay for this job?" are examples of this type of judgment.

Two prominent theories of cross-modality matching are *mapping* theory and *relation* theory (Krantz, 1972; Shepard, 1978). According to mapping theory, psychological values of stimuli from different continua are mapped onto a common scale of sensation and can be directly compared. A cross-modality match is presumed to occur when equal strength sensations are elicited by stimuli on different continua.

According to relation theory, relationships (e.g., ratios) between pairs of stimuli from different continua are compared. In physical measurement, a mass in grams cannot be compared with a length in centimeters but *ratios* of masses can be compared with ratios of length. By analogy, it may be possible to compare the ratio of the heaviness of two weights to the ratio of the loudness of two tones, since the ratios of stimulus pairs are on a common scale. Neither mapping theory nor relation theory gives an explicit account of contextual effects due to the stimulus distribution.

Another view of cross-modality matching, psychological *relativity* theory, contends that each stimulus is compared to its distribution, and the relative positions of the two stimuli with respect to their distributions are compared. For example, consider the question, "Is the weight as heavy as the light is bright?" Psychological relativity theory asserts that to answer this question, the relative position of the weight in the distribution of subjective heavinesses is compared to the relative position of the light in the distri-

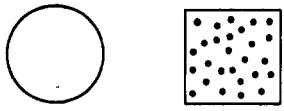

Figure 10. Example stimulus trial for the cross-modality experiment.

bution of brightnesses. The relative position of a stimulus in its distribution is assumed to be its range-frequency value, as in Equation 1.

Experiment 3 investigates the effects of the stimulus distribution (range and spacing) on cross-modality judgments. Two tasks were used, "difference" and "total intensity" judgments. Subjects were asked to compare and combine the subjective size of circles with the subjective darkness of dot patterns. Psychological relativity theory asserts that the estimated scale values derived from the subtractive and additive models (Equations 4 and 5) depend on the stimulus distribution for each dimension.

On the other hand, if the scale values are independent of the stimulus distribution, a simpler theory than relativity theory might suffice. A special case of Equations 4 and 5 in which the stimulus distribution only affects the judgment function (analogous to Equations 2b and 3b) can be written:

$$T_{ijk} = J'_k[c_j + s_i], \qquad (4b)$$

$$D_{ijk} = J_k[c_j - s_i], \qquad (5b)$$

where T and D are "total" and "difference" judgments, and the other terms are also as defined in Equations 4 and 5, except the scale values now do not depend on context (note there is no k subscript for the scale values). Experiment 3 examines whether the scale values in cross-modality judgment depend on stimulus distribution.

Method

Judges rated both "differences" and "total intensities" of the subjective size of circles and subjective darkness of dot patterns. An example stimulus trial is shown in Figure 10. Judges were asked to rate the "total intensity" of the size of the circle and the darkness of the dot pattern. They were also asked to indicate whether the size of the circle exceeded the darkness of the dot pattern and to estimate the "difference." Four different groups received the same set of circles paired with one of four distributions of dot patterns.

"Difference" task. Judges rated the "difference between the subjective size of the circle and the subjective darkness of the dot pattern" on a scale from −90 ("the darkness of the dot pattern is very very much greater than the size of the circle") to 90 ("the size of the circle is very very much greater than the darkness of the dot pattern"). Zero referred to an "equal match" of circle size and dot darkness.

"Total intensity" task. Subjects were instructed that the darker the dot pattern and the larger the circle, the greater the "total intensity" of the stimuli. "Total intensity" was rated on a scale from 0 ("the intensity is very very weak") to 90 ("the intensity is very very great"). On trials in which only one stimulus was presented, judges were asked to rate the "intensity" of that stimulus as though it was presented with another stimulus of zero "intensity." Instructions stated that the "intensity" of a stimulus presented alone was always less than the "total intensity" of the same stimulus presented with another.

Design. There were four experimental conditions. Six circles with diameters of 7.6, 11.2, 14.7, 18.3, 21.8, or 25.4 mm, were factorially combined with one of four different sets of dot patterns, with 1.5-mm black dots inside 25-mm squares. Six dot patterns were common to all four distributions of dots; these patterns contained 12, 18, 27, 40, 60 or 90 dots.

The four distributions of dots were as follows: Medium range: 10, 12, 18, 27, 40, 60, 90, and 135. There were 48 cells produced from a 6 × 8, Circle Size × Dot Number, factorial design. Wide range: 6, 12, 18, 27, 40, 60, 90, 180 (a 6 × 8 design). Positively skewed: 12, 14, 15, 16, 18, 21, 23, 27, 40, 60, 90 (a 6 × 11 design). Negatively skewed: 12, 18, 27, 40, 47, 51, 60, 70, 74, 77, 90 (a 6 × 11 design).

Each of the four conditions was constructed from a factorial design in which the stimulus on the left was one of six circles and the stimulus on the right was one of the eight or eleven dot patterns. Each circle and dot pattern was also judged by itself for the "total intensity" conditions.

Procedure. Each trial consisted of either a circle, a dot pattern, or both. Task order ("difference" or "total") was counterbalanced across subjects. Judges were given 24 to 26 representative warm-up trials to acquaint them with the stimulus range and stimulus spacing. Trials were printed in random order, and pages were shuffled to provide different orders. The task took approximately one hour.

Subjects. The judges were 157 undergraduates, with 38 to 41 different judges in each of the four conditions.

Results and Discussion

Figures 11 and 12 show mean judgments of the common stimuli in the "total" and "difference" tasks, respectively. Data points are connected by solid lines; dashed lines indicate prediction of the model described below. Estimated circle scale values are plotted on the abscissa with a separate curve for each level of the dot stimulus. The open cir-

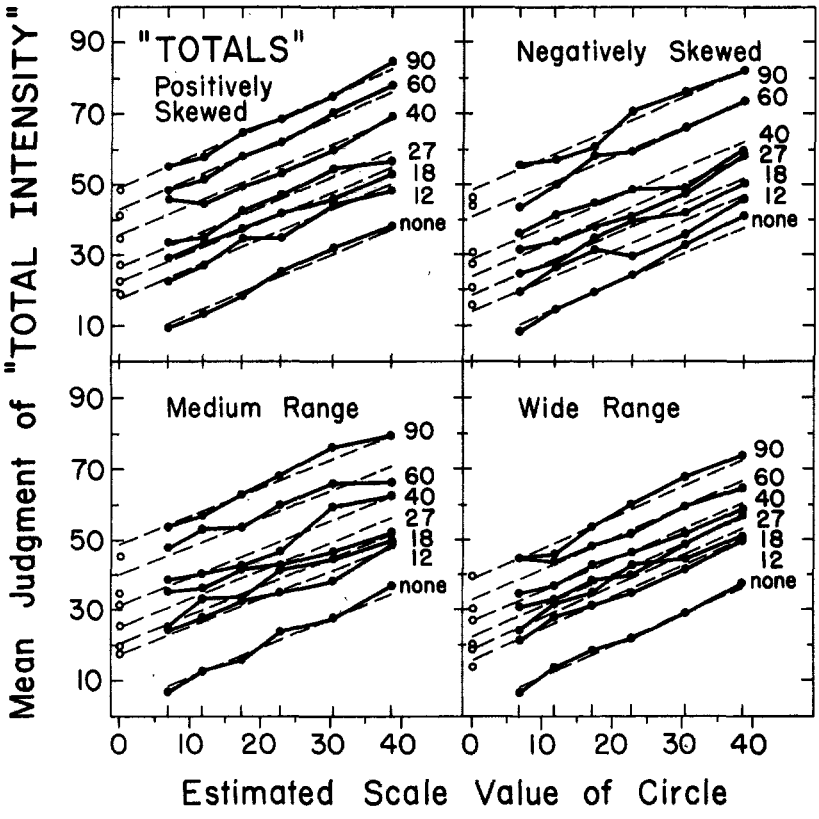

Figure 11. Mean "total intensity" judgments for common stimuli in the four contexts plotted as a function of the estimated circle scale value with a separate curve for each dot pattern. (Open circles are used for judgments of dot pattern alone; the curve labeled "none" show judgments of circles alone. Dashed lines are predictions of the additive model, in which scale values for dot patterns were assumed to depend on the range and spacing of the dot stimuli.)

cles in Figure 11 indicate judgments of dot stimuli alone; the bottom curves labeled "none" indicate judgments of circle stimuli alone. The average standard errors are 2.95 and 2.00 for the "difference" and "totals" tasks, respectively. The approximate parallelism of the curves in Figures 11 and 12 is consistent with predictions of additive and subtractive models for the "total" and "difference" tasks, respectively, assuming linear J functions.

The four sets of "total" judgments and four sets of "difference" judgments were fit to a special case of Equations 4 and 5:

$$\hat{T}_{ijk} = \alpha_k(c_j + s_{ik}) + \beta_k, \qquad (14)$$

$$\hat{D}_{ijk} = \gamma_k(c_j - s_{ik}), \qquad (15)$$

where $\hat{T}_{ijk}$ and $\hat{D}_{ijk}$ are the predicted "total intensity" and "difference" judgments of circle j and dot pattern i in context k; c_j and s_{ik} are the estimated scale values of the circles and the dot patterns, respectively; α_k and β_k, and γ_k are linear constants for each context. In Equation 15, when the response is "no difference" it is assumed that $c_j = s_{ik}$, that is, a cross-modality "match." In the additive model, the additive constant β_k is determined by the constraint that when a stimulus is not presented, its value is zero.[6]

Scale values for the dot patterns were permitted to be different for "totals" and "differences" in each context; that is, there are

[6] Equation 14 implies the following: $\hat{T}_{i0k} + \hat{T}_{0jk} - \hat{T}_{ijk} = \beta_k$, where $\hat{T}_{i0k}$ and $\hat{T}_{0jk}$ are predicted judgments of the ith dot pattern and jth circle presented alone in context k.

eight sets of scale values for dots. Since the distribution of circles was identical in all conditions, circle scale values were assumed to be the same for all eight conditions. By constraining the circle scale values to be identical across tasks, it is only necessary to fix one circle scale value in order to determine the darkness scale values.

The model shown in Equations 14 and 15 could account for all but 3.06% of the variance in the mean judgments. When it is assumed that the dot scale values are the same for both "difference" and "totals" tasks (but dot scales depend on context), the model left a residual of 4.41% using 38 fewer estimated parameters. Hence, not much is lost by assuming the dot scale values are largely independent of the task.

The vertical spacing of the curves in Figures 11 and 12 determines the scale values for the dots. Note that the spacing between the curves differs for different contexts. For example, the vertical spacing between the 12- and 90-dot curves is greater for the narrow range conditions (positive and negative skew) than for the wide range conditions for both "difference" and "total" tasks at all levels of circle size.

Estimated scale values are shown in Figure 13 as a function of log ϕ with separate curves for each stimulus distribution. Note that for both "differences" and "totals," the slopes are greater for the narrow range conditions than for the wide range condition. Note also that a medium-level dot pattern (e.g., 27) receives a greater scale value in the positively skewed context (where the majority of stimuli are lighter), than it does in the negatively skewed context. These changes in the slope and the height of the curves are in the general direction of the usual contextual effects in ratings (Parducci, 1974), although the curvature is less than expected by range-frequency theory.

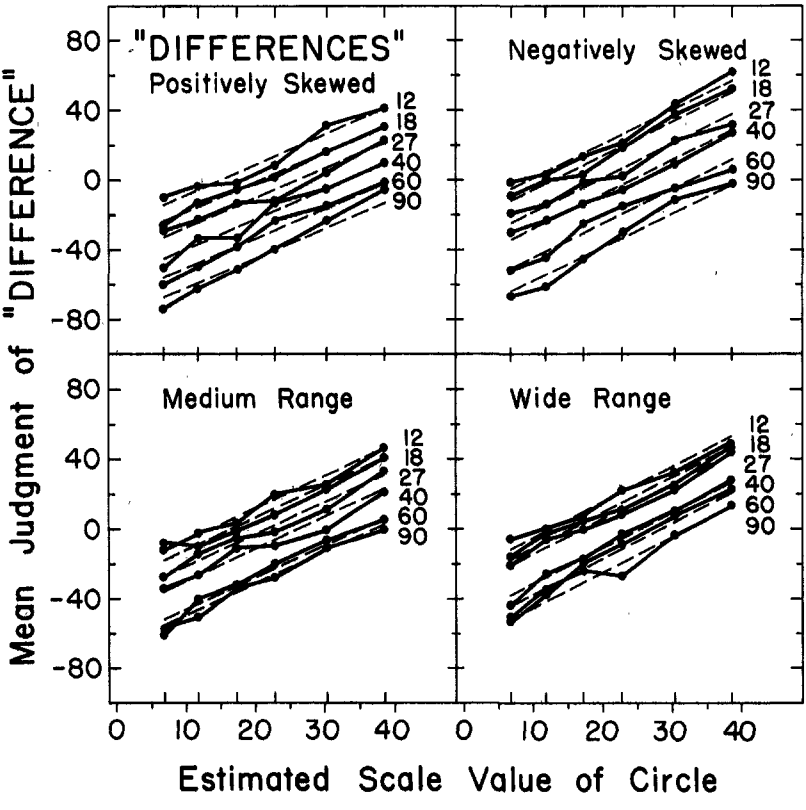

Figure 12. Mean "differences" for the common stimuli in the four tasks, plotted as in Figure 11. (Dashed lines are predictions of the subtractive model, allowing different dot scale values for different contexts.)

In summary, Experiment 3 shows that in cross-modality comparison and combination, the marginal stimulus distribution affects the scale values. Recall that in Experiment 2, the marginal stimulus spacing did *not* affect the scale values for within-modality comparisons. The results of Experiment 3 suggest that stimuli are judged within the context of other stimuli in the same continuum before they are compared or combined across continua. These findings are compatible with a psychological relativity view of cross-modality matching.

General Discussion

Single Stimulus Versus Comparison Judgments

Experiments 1 and 2 taken together suggest that contextual effects due to manipulation of the stimulus distribution for within-modal judgments can be attributed to the judgment function (J in Figure 1), rather than the psychophysical function. In Experiment 1, the stimulus distribution was shown to affect both category ratings and magnitude estimations of single stimuli. These contextual effects could be explained by a generalized form of Parducci's range-frequency theory for the judgment function. Localization of the contextual effects in J was consistent with the finding in Experiment 2 that subtractive model scale values do not appear to change as a function of stimulus spacing in "ratio" and "difference" judgments.[7]

Within-Modality Versus Cross-Modality Comparison

A comparison of Experiment 2 and 3 shows that manipulation of the stimulus spacing has different effects on the estimated scale values for within-modality and cross-modality judgments. In the within-modality judgments of Experiment 2, estimated scale values for "ratios" and "differences" in both contexts were virtually identical.

However, with cross-modality judgments of "differences" and "totals" in Experiment 3, estimated scale values *do* appear to depend on the range and spacing of the stimuli. Hence it appears that in cross-modality judgments, contextual effects occurred in the

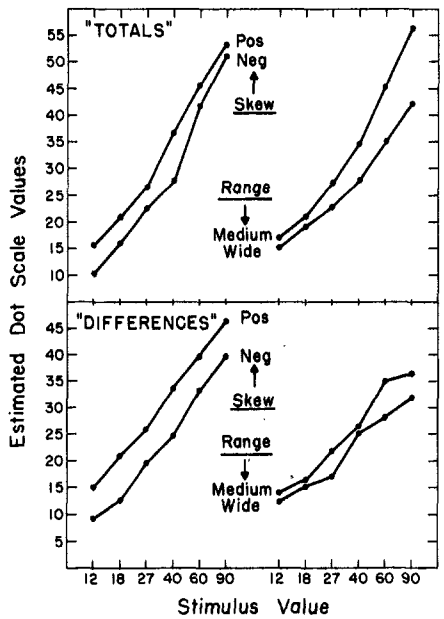

Figure 13. Estimated scale values for dot patterns plotted against stimulus values in equal log steps. (If there were no contextual effects, curves in each panel should coincide.)

H function; in within-modality judgments, the H function was unaffected. Cross-modality judgments may require an additional implicit judgmental transformation, in which subjective values are related to their own distributions before they are compared or combined across dimensions.

Some cross-modality judgments seem to involve a joint distribution between the two modalities in addition to the marginal distributions. Consider a judgment such as "Is this salary fair pay for this job?" Psychological relativity theory asserts that the relative position of the salary in the distribution of salaries is compared with the relative position of the job in the distribution of jobs. That is to say, judges compare the stimuli

[7] In Experiment 2, two types of distributions are relevant: (a) the marginal distribution, or spacing of the stimuli, and (b) the distribution of Ψ (subjective differences). It is assumed that the J function depends on the distribution of impressions (differences), which was not systematically manipulated in Experiment 2. Therefore, Experiment 2 was not designed to produce changes in the J function. Mellers and Birnbaum (in press) varied the J function by systematically manipulating the joint distribution in order to affect the distribution of Ψ (see also Birnbaum, et al, 1971, Experiment 5, and Mellers, 1982).

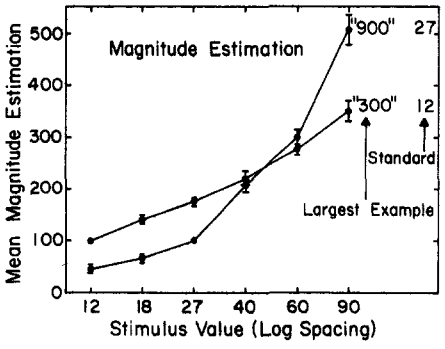

Figure 14. Evidence against between-subject comparison of magnitude estimation: Crossover refutes simple theories of magnitude estimation.

with respect to the marginal stimulus distributions. In addition to the marginal stimulus information, the judge may also attend to the distribution of salaries for each job and the distribution of jobs for each salary. Thus, the joint distribution seems relevant to certain cross-modality comparisons.

Mellers (1982, Experiment 4) investigated inequity judgments of hypothetical faculty members relative to other members of an academic department based on information about their merit ratings and salaries. The marginal distributions of merit and salary and the distribution of salary–merit differences were manipulated. It was shown that the marginal stimulus distributions affected the scale values as predicted by an extension of range-frequency theory, as in Experiment 3. Variations in the distribution of differences influenced the judgment function and could also be described by an extension of range-frequency theory applied to values of ψ, as in Birnbaum et al. (1971) and Mellers & Birnbaum (in press). In particular, Mellers found that a salary is judged to be "fair" when its position relative to the distribution of salaries corresponds to the relative position of the person's merit in the merit distribution. Furthermore, the judgment of inequity was found to depend on the subjective difference between salary and merit relative to the distribution of subjective differences.

Contextual Effects in Between-Subject Versus Within-Subject Designs

In a recent paper that examined "ratio" and "difference" judgments, different groups of subjects used different standards for the "ratio" task (Rule, Curtis, & Mullin, 1981). They assumed that such "ratio" judgments can be described by the following ratio model:

$$R_{ij} = J^*(s_j/s_i), \tag{16}$$

where R_{ij} is the estimation of "ratio," and s_j and s_i are the subjective values of the stimulus and the standard, respectively, and J^* is a single monotonic function that was assumed to be independent of the value of the standard. Violations of bilinearity in Figure 5 show that this model can be refuted if J^* is assumed to be any power function with an additive constant.

Even with the weak assumption that J* is monotonic, Equation 16 can be refuted. Figure 14 replots two of the curves from Figure 5 (Experiment 1) to reveal an ordinal violation of Equation 16 for between-subject designs. Figure 14 demonstrates that Equation 16 cannot account for the data when different subjects use different standards and different ranges of examples. For example, the "ratio" of the darkness of the 18-dot stimulus to the 12-dot stimulus (when the largest example is 300) is greater than the "ratio" of the 18-dot stimulus to the 27-dot stimulus (when the largest example is 900). In other words, the judgments and Equation 16 imply:

$$\frac{s_{18}}{s_{12}} > \frac{s_{18}}{s_{27}}. \tag{17}$$

However, the "ratio" of the 90-dot stimulus to the 27-dot stimulus (when the largest example is 900) is greater than the "ratio" of the 90-dot stimulus to the 12-dot stimulus (when the largest example is 300), implying:

$$\frac{s_{90}}{s_{27}} > \frac{s_{90}}{s_{12}}. \tag{18}$$

Thus, if Equation 16 is assumed, and J^* is the same for different groups of subjects, it follows that $s_{27} > s_{12}$ in Equation 17, but $s_{12} > s_{27}$ in Equation 18. This contradiction refutes Equation 16. However, these otherwise puzzling results are expected if different J^* functions occur for different groups of subjects, as in Equations 8 and 9, or 2b and 3b.

Another violation of Equation 16 can be seen in Figure 4. The "ratio" of the 27-dot

stimulus to the 12-dot stimulus when examples go up to 900 in the positively skewed context is greater than the "ratio" of the 40-dot stimulus to the 12-dot stimulus in the negatively skewed context, which shows that between-subject comparisons imply that the 27-dot pattern is darker than the 40-dot pattern! Yet *within* any of the conditions, the 40-dot stimulus is always judged to be darker than the 27-dot pattern. These results are consistent with Equations 8 and 9, which allow for different J^* functions.

Experiment 1 demonstrates that category ratings and magnitude estimations depend on the stimulus spacing and the response range given in the instructions. When different groups of subjects receive different stimulus or response distributions (i.e., magnitude estimations in which different groups of subjects receive different stimuli, standards, and/or different examples) the judgment function may indeed vary. Therefore magnitude estimations (or ratings) should not be regarded as an ordinal scale of sensation when compared across groups who experience different contexts.

If variations of standards and examples in between-subjects designs produce different J^* functions, the conclusion of Rule et al. (1981) that subjects use two operations when judging "differences" and "ratios" of heaviness of lifted weights should be questioned and reexamined. This conclusion rests entirely on the assumption that J^* in Equation 16 is the same for different groups of subjects who received different standards in their experiments. Figure 14 shows that this assumption may lead to contradictions.

It might be argued that Rule et al. (1981) were correct in their domain (heaviness) when they assumed that J^* was the same for all groups of subjects. However, it seems likely that if our Experiment 1 was replicated with lifted weights, similar contextual effects would occur in which the curves could be either made to fit the ratio model, give the same order as the subtractive model, or violate the ratio model depending on the response range and standard. The present theory (Equation 9) should be preferred to Equation 16 because it can explain both the Rule et al. results (which are by analogy replicated in Figure 5) and also our results (Figure 14), which contradict Equation 16.[8]

Is There a "Right" Way to Do Psychophysics?

Some have argued that with certain procedures, contextual effects can be avoided. For example, Poulton (1979), who reviewed contextual effects due to stimulus spacing and other factors, recommended using either a logarithmic spacing in magnitude estimation or a complete between-subjects design. According to range-frequency theory, logarithmic spacing could yield category ratings that are linearly related to subjective value if Fechner's law is assumed to be true. However, it seems unwise to assume Fechner's law in the design of experiments without providing a test of the assumption. Furthermore, for magnitude estimation, the judgment function appears to depend on the response examples. To "avoid" this issue, Poulton (1979) advised using no examples. However, when the examples are not manipulated, the J^* function is uncontrolled and unknown. Just because the experimenter refrained from using examples to illustrate the magnitude estimation scale does not mean that the subject did not do so. Each subject may use a different set of implicit examples, thereby producing a different J^* function.

If there are supposed to exist "right" procedures to avoid contextual effects, how are the "right" procedures to be determined? Some criteria for establishing the "correctness" of procedures are needed to resolve disagreements concerning proper procedure. Certainly, a procedure should not be advocated based on an uncertain theory that cannot be tested using that procedure (Birnbaum, 1982a; 1982b).

Consider Poulton's (1973; 1979) recommendation to "avoid" contextual effects by

[8] Some investigators have noted the difficulty of fitting the ratio model to data (Anderson, 1974; Eisler, 1960; Fagot & Stewart, 1971; Sjöberg, 1971). Ironically, Birnbaum and his colleagues, who argue for a subtractive rather than a ratio representation of "ratio" and "difference" judgments, have been reasonably successful in fitting the ratio model to "ratio" judgments. In their procedures, variation of standards and comparisons is done in within-subject factorial designs, and geometrically-spaced response examples are used. It is theorized that geometrically-spaced examples can induce an exponential J^* function for magnitude estimations. Hence, a subtractive comparison process can lead to data that will fit a ratio model (Birnbaum, 1980, 1982a).

asking each subject to make only a single judgment. From our theoretical viewpoint, this procedure confounds the stimulus and the context. If each subject judged a different "ratio," the judgment could be represented, $R_{ij} = J_{ij}(s_j - s_i)$, which shows that a different J_{ij} function would be allowed for each stimulus and standard. This recommendation must be based on the theory that each subject has the same J^* function as in Equation 16, a dubious assumption given the results of Figure 14. More importantly, such a procedure would not allow a test of the theory upon which it is based.

Another approach—the approach of this article—is to systematically manipulate the context and use a theory of the context to derive subjective values. Context is viewed as an integral part of the judgment process—something that cannot be "avoided." This *systextual design* approach is discussed in greater detail and compared with representative design, standardization design, and between-subjects design by Birnbaum (1974b; 1982a, Section E).

Conclusions

The following tentative conclusions appear consistent with the data:

1. Both category ratings and magnitude estimations appear to depend on the stimulus spacing and the response range in a similar fashion. These contextual effects are in the direction predicted by Parducci's range-frequency theory. The relationship between category ratings and magnitude estimations depends on the context and cannot be described by a context-invariant functional form.

2. In within-modal judgments, the scale values appear to be independent of variations in the stimulus distribution. These contextual effects can be accounted for by changes in the judgment function.

3. In cross-modality judgments, the scale values are influenced by the stimulus distribution. It appears that subjects compare the relative position of a stimulus in its distribution with the relative position of a stimulus of another modality to its distribution. Results were consistent with a psychological relativity theory of cross-modality judgment.

References

Anderson, N. H. Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press, 1974.

Anderson, N. H. On the role of context effects in psychophysical judgment. *Psychological Review*, 1975, *82*, 462–482.

Anderson, N. H. Cognitive algebra and social psychophysics. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement*. Hillsdale, N.J.: Erlbaum, 1982.

Birnbaum, M. H. The nonadditivity of personality impressions. *Journal of Experimental Psychology*, 1974, *102*, 543–561. (a)

Birnbaum. M. H. Using contextual effects to derive psychophysical scales. *Perception & Psychophysics*, 1974, *15*, 89–96. (b)

Birnbaum, M. H. Differences and ratios in psychological measurement. In N.J. Castellan & F. Restle (Eds.), *Cognitive Theory* (Vol. 3). Hillsdale, N.J.: Erlbaum, 1978.

Birnbaum, M. H. A comparison of two theories of "ratio" and "difference" judgments. *Journal of Experimental Psychology: General*, 1980, *3*, 304–319.

Birnbaum, M. H. Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement*. Hillsdale, N.J.: Erlbaum, 1982. (a)

Birnbaum, M. H. Problems with so-called "direct" scaling methods. In J. T. Kuznicki, R. A. Johnson, & A. F. Rutkiewic (Eds.), *Sensory methods: Problems and approaches to hedonics. ASTM STP773*. Philadelphia, Pa.: American Society for Testing and Materials, 1982. (b)

Birnbaum, M. H., Parducci, A., & Gifford, R. K. Contextual effects in information integration. *Journal of Experimental Psychology*, 1971, *88*, 158–170.

Chandler, J. D. Subroutine STEPIT-Finds local minima of a smooth function of several parameters. *Behavioral Science*, 1969, *14*, 81–82.

Eisler, H. Similarity in the continuum of heaviness with some methodological and theoretical considerations. *Scandinavian Journal of Psychology*, 1960, *1*, 69–81.

Eisler, H. Magnitude scales, category scales, and Fechnerian integration. *Psychological Review*, 1963, *70*, 243–253.

Fagot, R. F. Nested models of relative judgment: Applications to a similarity averaging model. *Perception & Psychophysics*, 1979, *26*, 255–264.

Fagot, R. F., & Stewart, M. Tests of product and additive scaling axioms. *Perception & Psychophysics*, 1971, *10*, 418–422.

Helson, H. *Adaption-level theory*. New York: Harper & Row, 1964.

Johnson, D. M., & Mullally, C. R. Correlation-and-regression model for category judgments. *Psychological Review*, 1969, *76*, 205–215.

Krantz, D. H. Magnitude estimations and cross-modality matching. *Journal of Mathematical Psychology*, 1972, *9*, 168–199.

Marks, L. E. On scales of sensation: Prolegomena to any future psychophysics that will be able to come

forth as a science. *Perception & Psychophysics*, 1974, *16*, 358–376.

Marks, L. E. A theory of loudness and loudness judgments. *Psychological Review*, 1979, *86*, 256–285.

Mellers, B. A. Equity judgment: A revision of Aristotelian views. *Journal of Experimental Psychology: General*, 1982, *111*, 242–270.

Mellers, B. A. & Birnbaum, M. H. Contextual effects in social judgment. *Journal of Experimental Social Psychology*, in press.

Montgomery, H. Direct estimation: Effect of methodological factors on scale type. *Scandinavian Journal of Psychology*, 1975, *16*, 19–29.

Moskowitz, H. R. Utilitarian benefits of magnitude estimation scaling for testing product acceptability. In J. T. Kuznicki, R. A. Johnson, and A. F. Rutkiewic (Eds.), *Selected sensory methods: Problems and approaches to hedonics. ASTM STP 773*. Philadelphia, Pa.: American Society for Testing and Materials, 1982.

Parducci, A. Range-frequency compromise in judgment. *Psychological Monographs*, 1963, *77*(2, Whole No. 565).

Parducci, A. Category judgment: A range-frequency model. *Psychological Review*, 1965, *72*, 407–418.

Parducci, A. The relativism of absolute judgment. *Scientific American*, 1968, *219*, 84–90.

Parducci, A. Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of Perception* (Vol. 2). New York: Academic Press, 1974.

Parducci, A. Category ratings: Still more contextual effects. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement*. Hillsdale, N.J.: Erlbaum, 1982.

Parducci, A., & Perrett, L. Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology*, 1971, *89*, 427–452.

Poulton, E. C. The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, 1968, *69*, 1–19.

Poulton, E. C. Unwanted range effects from using within-subject experimental design. *Psychological Bulletin*, 1973, *80*, 113–121.

Poulton, E. C. Models for biases in judging sensory magnitude. *Psychological Bulletin*, 1979, *86*, 777–803.

Restle, F., & Greeno, J. G. *Introduction to mathematical psychology*. Reading, Mass.: Addison-Wesley, 1970.

Robinson, G. H. Biasing power law exponents in magnitude estimation instructions. *Perception & Psychophysics*, 1976, *19*, 80–84.

Rule, S. J., Curtis, D. W., & Mullin, L. C. Subjective ratios and differences in perceived heaviness. *Journal of Experimental Psychology: Human Perception and Performance*, 1981, *7*, 459–466.

Shepard, R. N. On the status of "direct" psychological measurement. In C. W. Savage (Ed.), *Minnesota studies in the philosophy of science* (Vol. 9). Minneapolis: University of Minnesota Press, 1978.

Sjöberg, L. Three models for the analysis of subjective ratios. *Scandinavian Journal of Psychology*, 1971, *12*, 217–240.

Stevens, S. S. A metric for the social consensus. *Science*, 1966, *151*, 530–541.

Stevens, S. S. Issues in psychophysical measurement. *Psychological Review*, 1971, *78*, 426–450.

Stevens, S. S., & Galanter, E. H. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, *54*, 337–411.

Teghtsoonian, R. On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review*, 1971, *78*, 71–80.

Torgerson, W. S. Distances and ratios in psychological scaling. *Acta Psychologica*, 1961, *19*, 201–205.

Received August 12, 1981 ■