Journal of Experimental Psychology:
Human Perception and Performance
1990, Vol. 16, No. 3, 613–625

0096-1523/90/$00.75

Judgments of Proportions

Carol A. Varey
University of California, Berkeley

Barbara A. Mellers
University of California, Berkeley

Michael H. Birnbaum
California State University, Fullerton

This study investigated the processes that underlie estimates of relative frequency. Ss performed 4 tasks using the same stimuli (squares containing black and white dots); they judged "percentages" of white dots, "percentages" of black dots, "ratios" of black dots to white dots, and "differences" between the number of black and white dots. Results were consistent with the theory that Ss used the instructed operations with the same scale values in all tasks. Despite the use of the correct operation, Ss consistently overestimated small proportions and underestimated large proportions. Variations in the distributions of actual proportions affected the extent to which Ss overestimated small proportions and underestimated large proportions in the direction predicted by range-frequency theory. Results suggest that proportion judgments, and by analogy probability judgments, should not be taken at face value.

Many of our real-world decisions are based on subjective probabilities. Whether we bring an umbrella to work depends on our estimate of the chance of rain; whether we buy a lottery ticket depends, in part, on our estimate of the probability of winning; whether we support nuclear power is influenced by our beliefs about the likelihood of disastrous accidents.

Subjective probabilities arise from a complex mixture of our perceptions, memories, and reasoning processes. For example, to estimate the probability of rain, we might take into consideration such information as the appearance of the sky, our knowledge of past weather conditions in the area, recent weather reports, and our opinions of the weather forecasters. How such information is retrieved from memory, evaluated, and combined to form an estimate of subjective probability has been the focus of much research (e.g., Birnbaum, 1983; Birnbaum & Mellers, 1983; Edwards, 1968; Hasher & Zacks, 1984; Kahneman, Slovic, & Tversky, 1982; Peterson & Beach, 1967; Schum, 1981; Shanteau, 1974; Slovic, Lichtenstein, & Fischhoff, 1988; Wallsten, Budescu, Rappoport, Zwick, & Forsyth, 1986; Zadeh, 1975).

At the heart of probability estimation is the concept of relative frequency or proportion. The present article examines judgments of relative frequency by using simple situations in which all of the necessary information is available. The first experiment investigates operations that underlie percentage judgments. The second experiment tests the hypothesis that contextual effects influence percentage judgments.

Experiment 1

Early work on proportion judgments investigated the relationship between judged proportion and actual proportion. Visual displays containing two types of elements were presented to subjects, who were asked to estimate the proportion of elements of one type. Philip (1947) presented subjects with 11 stimuli (cards containing two colors of dots in proportions ranging from 13/36 to 23/36) and obtained judgments of proportion on an 11-point rating scale. Philip found a linear relationship between ratings and actual proportions.

Stevens and Galanter (1957) used a wider range of stimuli (11 cards containing two colors of dots with proportions ranging from 3/36 to 33/36) and found that both 7-point category ratings and percentage estimates were related to actual proportions by inverse ogival-shaped functions (slopes of the curves were flatter in the center and steeper at the ends).

Shuford (1961) used visual displays containing 400 elements (vertical and horizontal lines) and asked subjects to judge proportions directly on a scale from 0% to 100%. By using actual proportions ranging from 40/400 to 360/400 at 10% intervals, Shuford found a linear relationship between judged percentages and actual percentages. In summary, the experiments found different stimulus–response functions, depending on the experimental conditions. However, these studies are open to numerous alternative interpretations when it is conceded that the output (judgment) function as well as the psychophysical function affects responses.

The present article extends this earlier work in two ways. First, relative frequency and total frequency are unconfounded to allow tests of alternative models. Subjects were shown squares containing black and white dots in which the number of black and white dots were independently varied in a factorial design. With this design, the same physical proportion is constructed from different numbers of dots. Second,

Support for this research was provided by a National Science Foundation grant to Barbara A. Mellers (BNS-845-1368).

Correspondence concerning this article should be addressed to Barbara A. Mellers, Department of Psychology, University of California, Berkeley, California 94720.

subjects performed four tasks using the same stimuli: "differences," "ratios," "percentage white," and "percentage black."[1] By obtaining judgments from different tasks, greater leverage is provided to distinguish among alternative representations of the underlying operations and judgment functions (or the output mappings from impressions to responses). If "percentage" judgments are obtained in isolation and if they are ordinally consistent with a relative ratio (or percentage) operation, they cannot be distinguished on ordinal grounds from a ratio model or a subtractive model. In this case, there would be no ordinal grounds for selecting one model over another. However, if "percentage" judgments are obtained in conjunction with "difference" judgments and "ratio" judgments and if the scale convergence criterion is assumed (Birnbaum, 1974, 1982; Birnbaum & Veit, 1974), it becomes possible to rule out sets of models for the four tasks in favor of other sets of models.

The scale convergence criterion is the premise that the subjective values of the stimuli are independent of the task. If scale convergence is assumed and different operations underlie the judgments, then the rank orders of the judgments across the different tasks should not be monotonically related but instead should have certain predictable patterns. Two theories that assume the scale convergence criterion are discussed next.

One-Operation Theory

One possibility is that people do not have the "mental machinery" to perform different operations. Regardless of the instructions; they might use a single operation and map their impressions to the appropriate response scale. For example, if subjects use a subtractive operation in all four tasks and if the subjective values of the stimuli are independent of the task, then all four sets of judgments would be monotonically related. Another version of the one-operation theory might be that subjects use a ratio operation in all four tasks and then transform their subjective ratios to the appropriate response scale. One-operation theory implies that the rank order of judgments will be the same across all four tasks. Birnbaum (1978) used the term *indeterminacy theory* for this case because if there is only one operation, it is not possible to determine whether that operation is a ratio or a difference.

Multiple-Operations Theory

Birnbaum and his colleagues (see Birnbaum, 1978, 1980, 1982, for reviews) distinguished between the following two types of situations. For many continua, when people are asked to judge "ratios" and "differences" of stimuli, they appear to use a subtractive operation regardless of the instructions. Judgments of "ratios" and "differences" are monotonically related for continua such as loudness and pitch of tones (Birnbaum & Elmasian, 1977; Elmasian & Birnbaum, 1984), heaviness of weights (Mellers, Davis, & Birnbaum, 1984), likableness of persons described by trait adjectives, darkness of dot patterns, and many others. However, when subjects are asked to judge "ratios" and "differences" of distances between stimuli, they appear to use both ratio and subtractive operations as instructed (Birnbaum, 1978, 1982; Birnbaum, Anderson, & Hynan, 1989; Veit, 1978).

These results have been interpreted as follows: When the stimuli along a subjective continuum form an interval scale, the ratio operation is meaningless and subjects compare stimuli by means of subtraction. However, even on an interval scale, "distances" or "differences" have a well-defined zero point; hence, "ratios of differences" and "ratios of distances" are meaningful. In this case, subjects can perform either ratio or difference operations. For example, when subjects are asked to judge "ratios" and "differences" of easterliness and westerliness of U.S. cities, all four sets of judgments are monotonically related (Birnbaum & Mellers, 1978), consistent with a one-operation theory. However, when instructed to judge "ratios" and "differences" of distances between pairs of the same U.S. cities, subjects produce two different rank orders, consistent with the operations on a ratio scale of distances (Birnbaum et al., 1989). Scale values derived from the fit of the instructed operations to the data converge with scale values obtained from the fit of the subtractive operation to simple "ratio" and "difference" judgments.

In the present tasks, subjective values associated with the number of white or black dots might correspond to distances or differences between the physical number of dots and zero dots. If the stimuli themselves are inherently subjective distances or differences, subjects could meaningfully perform both ratio and difference operations. However, if the stimuli are merely points along a continuum with an undefined zero point, subjects might be expected to perform a subtractive operation in all four tasks, regardless of the instructions.

Method

Stimuli consisted of squares containing different numbers of black and white dots. Each subject judged the "difference" between the numbers of black and white dots, the "ratio" of the number of black dots to white dots, the "percentage" of white dots, and the "percentage" of black dots.

Stimuli and design. Subjects were presented with 4-cm squares containing different numbers of black and white dots. An example is shown in Figure 1. Thirty-six stimuli were constructed from a 6 × 6 (White Dot × Black Dot) factorial design. Numbers of white dots and black dots were approximately equally spaced on a logarithmic scale. White dot frequencies were 5, 8, 12, 18, 27, and 40; black dot frequencies were 40, 60, 90, 135, 201, and 301. These levels unconfound total frequencies from relative frequencies, as shown in Table 1. Entries are physical percentages that remain approximately constant along the diagonals despite different total number of dots.

Instructions. Subjects read a general instruction sheet that described the stimuli and the four tasks. Specific instructions for each

[1] Quotation marks are used to distinguish instructions and judgments from the theoretical operations that subjects might use when making their judgments. For example, when the instruction is to judge "ratios," the ratio model may or may not underlie "ratio" judgments.

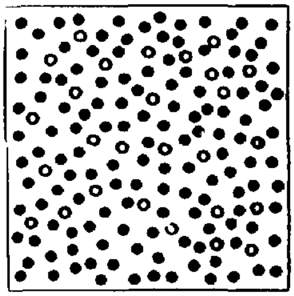

Figure 1. An example stimulus with 27 white (open) dots and 135 black (filled) dots (17% white and 83% black).

task followed, in which subjects were asked to make intuitive judgments of the "percentage" of black dots, the "percentage" of white dots, the "ratio" of black dots to white dots, and the "difference" between the number of black dots and white dots. Subjects were shown an example stimulus that contained two white and two black dots. Subjects were told that for this stimulus, their responses should be 50% black dots, 50% white dots, a ratio of 1, and a difference of 0.

There are four main differences between the procedures used in the present studies and those used in previous research on "ratios" and "differences" (Birnbaum, 1978, 1980). First, in the "difference" task, subjects were asked to estimate physical differences rather than make ratings on a category scale. Subjects were told to estimate the actual numerical difference between the number of black dots and white dots. Second, in the "ratio" task, the modulus was set to 1.0 rather than 100. Third, there was only one example stimulus (and example response) presented in the instructions. Fourth, differences presented for judgment were always nonnegative, because the number of black dots was never less than the number of white dots.

Procedure. Stimuli were arranged in five-page booklets. The first page contained six stimuli (selected from the 36 trials) that served as practice trials. The following four pages contained the 36 experimental stimuli presented in random order. Page order for the experimental trials was counterbalanced by using two Latin square designs.

Task order and page order for the booklets were counterbalanced as follows: Four different task orders were crossed with the four booklet orders to form 16 possible order combinations. There were 3 subjects in each ordering. Subjects were tested 2 to 8 at a time and worked alone at their own pace. The experiment took approximately 1 hr.

Participants. Forty-eight undergraduates at the University of California, Berkeley, received credit in an introductory psychology course for participating. A few additional subjects who failed to complete the tasks in the allotted time were excluded from the analyses.

Table 1
Physical Proportions of White Dots

	White dot frequencies					
Black dot frequencies	5	8	12	18	27	40
40	.111	.167	.230	.310	.403	.500
60	.077	.112	.167	.231	.310	.400
90	.053	.082	.118	.167	.231	.308
135	.036	.056	.082	.118	.167	.229
201	.024	.038	.056	.082	.118	.166
301	.016	.026	.038	.056	.082	.117

Results

Figure 2 shows mean responses in the four tasks plotted as a function of the estimated scale values for white dots, with a separate symbol for each level of black dots. (Lines represent predictions of a theory that will be discussed later.) Panel A shows judged "differences." If the data were consistent with a subtractive model and if the response function was linear, curves connecting data symbols would be parallel. Although the interaction between black and white dots was statistically significant,[2] $F(25, 1175) = 6.31$, deviations from parallelism do not appear to conform to any interpretable pattern.

Panel B presents judged "ratios" as a function of estimated scale values for the number of white dots (in reverse order on the abscissa). If subjects used a ratio operation when making their judgments and if the response scale was linear, the curves connecting data points would diverge (differences between the curves would increase from left to right). The data approximate this predicted pattern of divergence. The interaction between black and white dots was statistically significant, $F(25, 1175) = 14.51$. To test whether this divergence was representative of individual data, vertical differences in responses between the two most extreme curves (the 301 and 40 black dots curves) were computed for 5 white dots (rightmost points) and 40 white dots (leftmost points). The majority of subjects (87%) had a larger difference at 5 white dots than at 40, consistent with the pattern of the means.

Mean "percentage white" and "percentage black" judgments are presented in Panels C and D of Figure 2, respectively. According to the relative ratio model, the curves in both panels should be closer together at one end and bulge out at the other end (resembling two different ends of a football). The data in both panels resemble these trends. The interactions between white and black dots were significant: $F(25, 1175) = 17.86$ and $F(25, 1175) = 9.63$ for "percentage white" and "percentage black" judgments, respectively. Differences in responses between the two most extreme curves were again computed for 5 and 40 white dots. For "percentage white" and "percentage black" judgments, 94% and 83% of the subjects, respectively, showed the same pattern of divergence or convergence as the means.

Ordinal relationships among the tasks. Figure 3 plots the mean responses in each of the tasks against those of the other tasks, with a separate panel for each of the six possible combinations of four tasks and a separate symbol for each number of black dots. The top panels show judged "differences" plotted against "ratios" (Panel A), "percentage black" (Panel B), and "percentage white" (Panel C). In all three of the top panels, no single monotonic function can describe the relationship between judged "differences" and the other tasks; judged "differences" are ordinally distinct from judgments of "ratios," "percentage black," and "percentage white."

The lower three panels (D, E, and F) show the relationships between "ratios" and "percentage black" (Panel D), "ratios" and "percentage white" (panel E), and "percentage black" and

[2] The word *significant* is used throughout this article to denote $p < .01$.

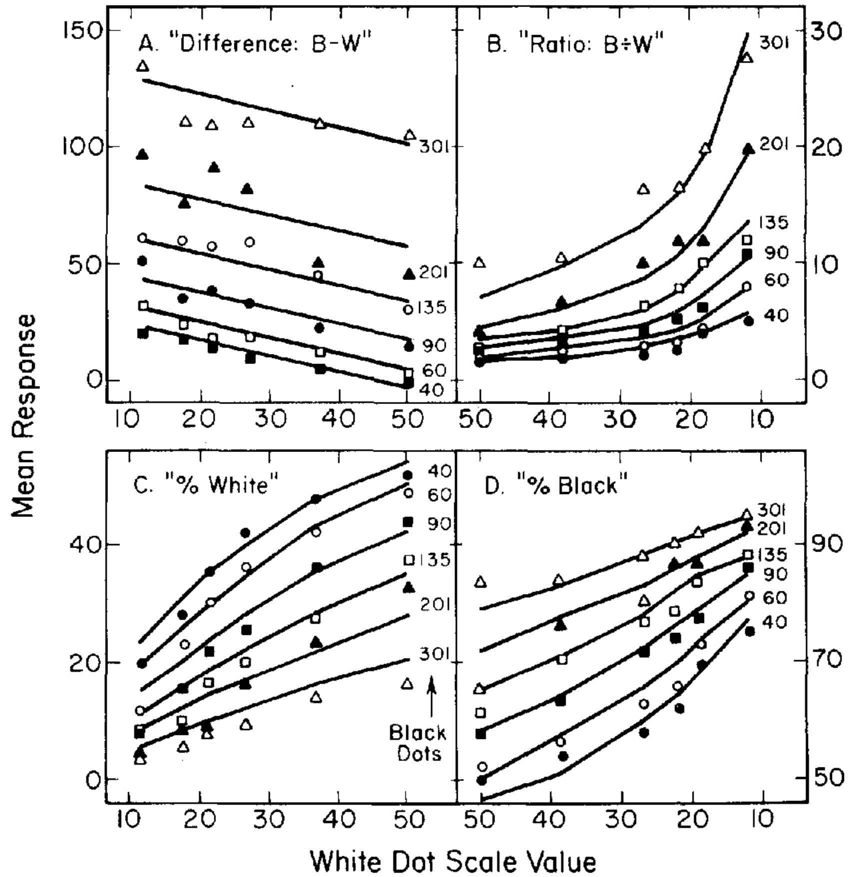

Figure 2. Mean responses in the four tasks plotted as a function of the estimated scale values for white (W) dots, with a separate curve and symbol for each number of black (B) dots. (Panel A presents the "difference" judgments, Panel B shows "ratio" judgments, Panel C shows "percentage white" judgments, and Panel D shows "percentage black" judgments. Lines show predictions of the theory that subjects use the instructed operation with the same scale values.)

"percentage white" (Panel F). In Panels D and F, the points seem to cluster about a single curve, whereas in Panel E, there are some deviations from a single monotonic function, but those deviations do not appear to be large.

The six panels in Figure 3 are inconsistent with the hypothesis that a single operation underlies all four sets of judgments. One-operation theory implies that judgments from each pair of tasks should be monotonically related. Instead, the data show that the rank order of "differences" is systematically different from the other three rank orders, as predicted by the theory that subjects used the instructed operation in all four tasks.

Fit of the multiple-operation theory. Mean responses in all four tasks were simultaneously fit to the theory that subjects perform the tasks as instructed and that the subjective values of black and white dots are identical across all four tasks. This theory can be expressed in the following four equations:

$$D_{ij} = a_D(b_j - w_i) + c_D \quad (1)$$

$$R_{ij} = a_R(b_j/w_i) \quad (2)$$

$$W_{ij} = 100w_i/(b_j + w_i) \quad (3)$$

$$B_{ij} = 100b_j/(b_j + w_i)\,, \quad (4)$$

where D_{ij}, R_{ij}, W_{ij}, and B_{ij} refer to predictions for "differences," "ratios," "percentage white," and "percentage black," respectively; w_i and b_j are the subjective values of white and black dots; and a_D, a_R, and c_D are linear constants for "differences" and "ratios."

For each task, a proportion of variance of deviations from the model was defined as follows:

$$P_T = \Sigma\Sigma(T_{ij} - \hat{T}_{ij})^2/\Sigma\Sigma(T_{ij} - \bar{T})^2\,, \quad (5)$$

where P_T is the proportion of systematic residual variance for each task, T; T_{ij} is the mean judgment for cell ij; $\hat{T}_{ik}$ is the corresponding prediction; and $\bar{T}$ is the overall mean judgment for task T. A computer program was written to select parameters to minimize the sum of these four proportions of deviations, using Chandler's (1969) STEPIT subroutine to accomplish the minimization. Fourteen parameters were estimated from the 4 × 36 data cells; there were six scale values for white dots and five scale values for black dots, with the smallest black scale value fixed to its physical value (40 dots), and three linear constants, as shown in Equations 1–4.

The proportion of systematic residual variance (P_T) in the four designs was .037, .032, .047, and .034 for "differences," "ratios," "percentage white," and "percentage black," respec-

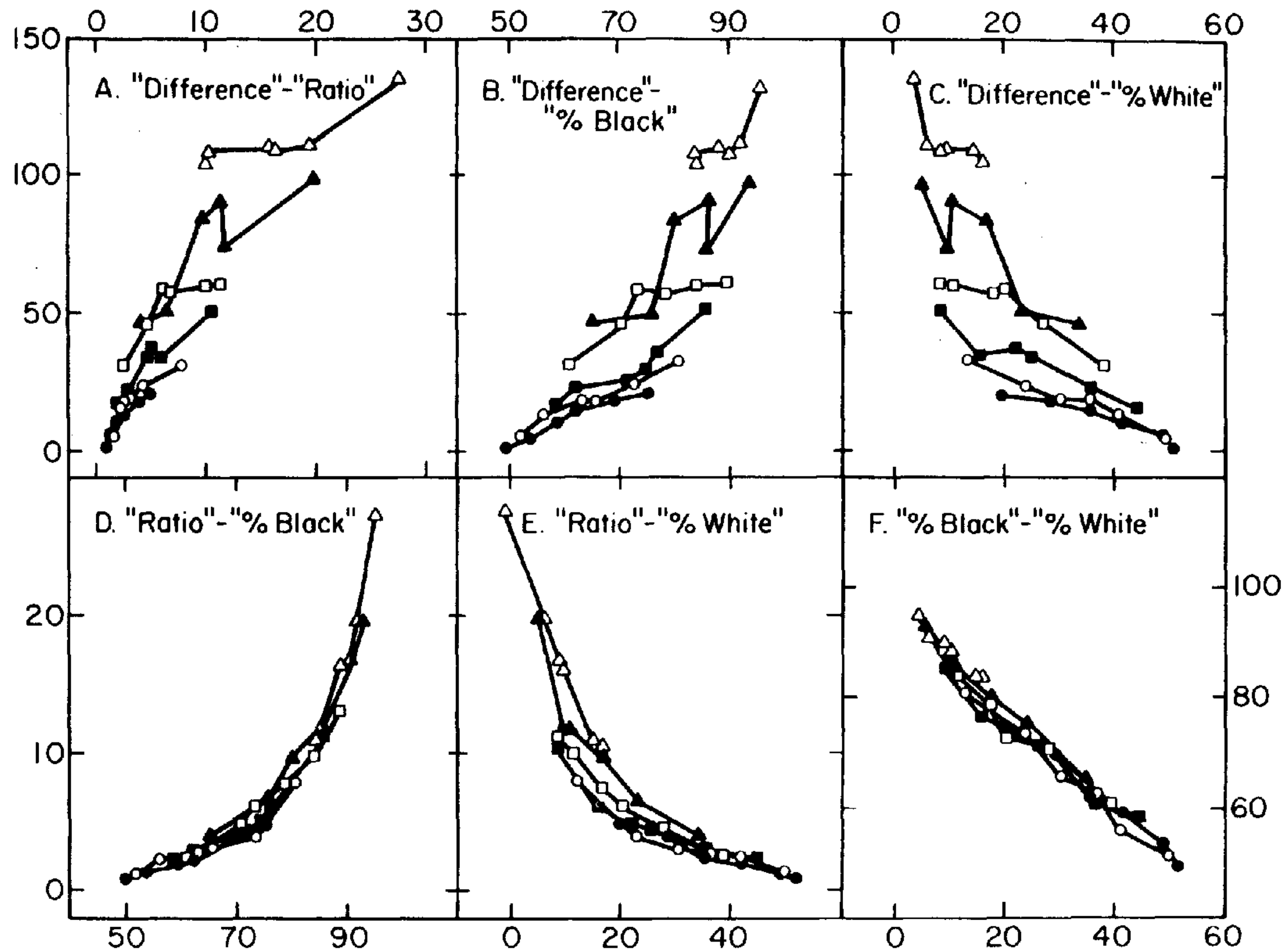

Figure 3. Mean responses from each task plotted against the other tasks, with a separate symbol for each number of black dots and a separate point for each of the common stimuli. (Lines connect stimuli with the same number of black dots. Upper panels show the "difference" judgments plotted against "ratios" [Panel A], black "percentages" [Panel B], and white "percentages" [Panel C]. Lower panels show "ratios" plotted against black "percentages" [Panel D], "ratios" plotted against white "percentages" [Panel E], and black against white "percentages" [Panel F].)

tively. The fit of the theory can be assessed in Figure 2, which shows the predictions (solid lines) plotted with the data (symbols).

Figure 4 presents the estimated scale values for white and black dots (w_i and b_j) plotted against physical values. The subjective values for white dots (open circles) are larger than the physical values, whereas scale values for black dots (solid circles) are smaller than their physical values (with the exception of 40 black dots, which was set to its physical value). The relationship between estimated scales and physical scales for black dots and white dots may be due to the fact that the actual size of the white dots was slightly larger than that of the black dots. It is also possible that white dot scale values were larger than their physical values because there were always fewer white dots than black dots or because of asymmetries in perception, such as the "pop-out" effects investigated by Treisman (1988).

The estimated scale values for numbers of dots in Figure 4 can be considered in light of previous results from experiments that investigated judgments of frequency by using latency or threshold accuracy measures (Beckwith & Restle, 1966; Jensen, Reese, & Reese, 1950; Jevons, 1871; Kaufman, Lord, Reese, & Volkmann, 1949). Estimates of number are usually accurate for small numbers and fall short as the number increases. This research led investigators to postulate three processes for judging frequency (Klahr, 1973; Klahr & Wallace, 1973). For fewer than 5 items, they posit a "subitizing" process; for between 5 and 20 items, they assume a "counting" process; and for more than 20 items, they propose an "estimation" process. Although the points in Figure 4 are scale values rather than estimates of number, they show negative acceleration, compatible with previous findings for frequency estimations.

Judged proportion versus actual proportion. Figure 5 plots mean judgments from the "percentage white" task (open circles) and the "percentage black" task (solid circles) averaged over stimuli with equal physical proportions. If judged percentages and actual percentages were identical, the points

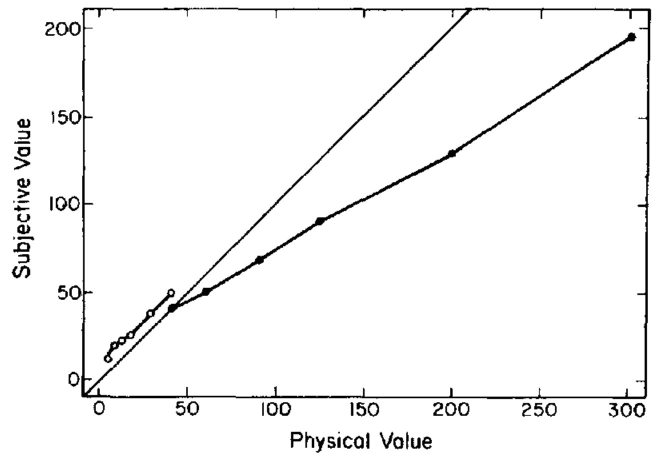

Figure 4. Estimated scale values for white dots (open circles) and black dots (solid circles) plotted against physical values.

would fall on the diagonal (identity) line. However, judged percentages below 50% are overestimated, and those above 50% are underestimated. To investigate this effect for individuals, the average judged percentage for each subject was compared with the average physical percentage in each task; 98% of the subjects both overestimated in "percentage white" judgments and underestimated in "percentage black" judgments.

The curves in Figure 5 display the predictions of the multiple-operations theory averaged over stimuli with equal physical proportions. As seen in the figure, the theory captures the pattern of overestimation and underestimation of judged percentages, although it seems to exaggerate the difference between black and white dots at 50%.

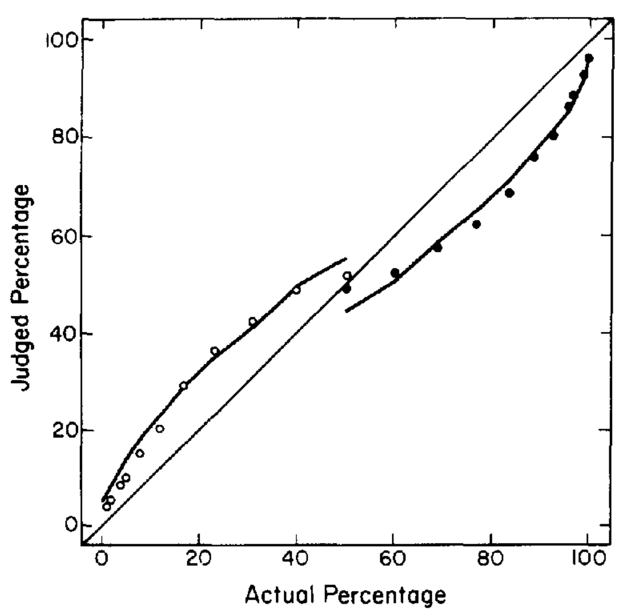

Figure 5. Judged "percentages" (averaged over equal physical percentage) plotted against actual percentages. (White "percentages" and black "percentages" are shown as open and solid circles, respectively. Lines show best-fit predictions of the theory discussed in the text.)

The overestimation of small percentages and underestimation of larger ones follows from the relative ratio operation and psychophysical functions having the property that the ratio of the scale value associated with the smaller number of elements (in this case, white dots) to the scale value for the larger number of elements (black dots) is greater than the corresponding physical ratio (i.e., $w/b > \varphi_w/\varphi_b$). This condition is satisfied by the "regressed" function in Figure 4, which has a slope less than one. Several other psychophysical functions also have this property (e.g., power functions with exponents less than one) and would also imply this pattern of overestimation and underestimation. Thus, even thou h people are assumed to be using the "correct" operation, inaccurate judgments can be explained as a consequence of the shape of the psychophysical functions.

Experiment 2

The purpose of Experiment 2 was to investigate changes in "percentage" judgments that are due to variations in the stimulus distribution. For many continua, it has been found that the same stimulus receives different judgments depending on the distribution of other stimuli presented for judgment. When larger stimuli are presented with greater frequency, an intermediate-valued stimulus will typically receive a smaller judgment. When smaller stimuli are presented with greater frequency, the same stimulus will typically receive a larger judgment. These contextual effects have been observed for category ratings, magnitude estimations, "absolute" number estimates, and other numerical responses (Mellers, 1986; Mellers & Birnbaum, 1982, 1983; Parducci & Perrett, 1971). Thus, it is possible that proportion judgments will show similar effects caused by changes in the stimulus distribution.

However, it may be that proportion judgments are independent of the context, because of the nature of the stimulus and response scales. For example, when subjects judge the proportion of elements of one color in a two-color array, the scale might have natural anchors at 0 (when there are no members of one color), at 0.5 (when the two are equal), and at 1 (when there are no members of the other color). These constraints might suffice to pin down the response scale and therefore preclude or substantially reduce the possibility of contextual effects in proportion judgments.

Method

There were four conditions in Experiment 2, constructed from a 2 × 2 (Task × Context) factorial design, with different subjects in each condition. Stimuli were squares containing black and white dots, as in Experiment 1, and subjects' tasks were to judge either the "percentage" of white dots or the "percentage" of black dots. The distribution of percentages (context) was either positively skewed or negatively skewed.

Stimuli and design. Each of the four conditions contained 150 stimuli. There were 130 contextual stimuli and 20 experimental stimuli that were common to all four conditions. The common stimuli were a subset of the design from Experiment 1 and constituted a 4 × 5 (White Dot × Black Dot) factorial design. Numbers of white dots were 8, 12, 18, and 27; numbers of black dots were 40, 60, 90, 135, and 201.

Figure 6 shows how the 130 contextual stimuli were distributed in relation to the stimuli in the 4 × 5 common design. Because it was thought that percentage judgments based on arrays containing two types of elements might be anchored at three points—0%, 50%, and 100%—the contextual manipulations were carried out separately on percentages ranging from 0 to 50 (white percentages) and from 50 to 100 (black percentages).

In the upper panels of Figure 6, actual percentages for the common design are plotted as a function of the number of white dots, with a separate curve and symbol for each number of black dots. The numbers in Panel A show the frequency and approximate value of the black and white dots for contextual stimuli in the positively skewed distribution. For example, 100 in Panel A indicates that there were 100 stimuli having from 9 to 11 white dots and from 136 to 200 black dots (percentages of white ranged from 3.7% to 5.5%). In this condition, the distribution of white percentages is positively skewed, and the marginal distributions of white and black dots are positively and negatively skewed, respectively. The numbers in Panel B show the frequency and values of contextual stimuli for the other context. In this condition, the distribution of white percentages is negatively skewed, and the marginal distributions of white and black dots are negatively and positively skewed, respectively.

Lower panels show the physical percentages of black dots from the common designs. Panel C depicts the positively skewed distribution of black percentages, which arises from the same stimuli as in Panel B; a negatively skewed distribution of white percentages produces a positively skewed distribution of black percentages. Similarly, Panel D arises from the same distribution as in Panel A; white percentages that are negatively skewed produce black percentages that are positively skewed.

Instructions. Subjects were asked to judge either the "percentage" of white dots or the "percentage" of black dots in each stimulus. As in Experiment 1, instructions included a stimulus with two white dots and two black dots. Subjects were told that for this stimulus, they should respond 50%. Instructions emphasized that the judgments should be subjective rather than based on actual computations.

Procedure. The stimuli were arranged in 15-page booklets, with 10 stimuli on each page. To acquaint the subjects with the full range of proportions, we presented on the first page eight contextual stimuli and the smallest and largest percentages among the common stimuli. The next 5 pages contained only contextual stimuli. Each of the remaining 10 pages contained eight contextual stimuli and two common stimuli. Ten different versions of the booklet were created by ordering the last 10 pages according to a Latin square design. Subjects were tested 1 to 5 at a time and worked alone at their own pace. The experiment took approximately 1 hr.

Participants. Subjects were 158 undergraduates from the University of California, Berkeley, who participated for course credit in an

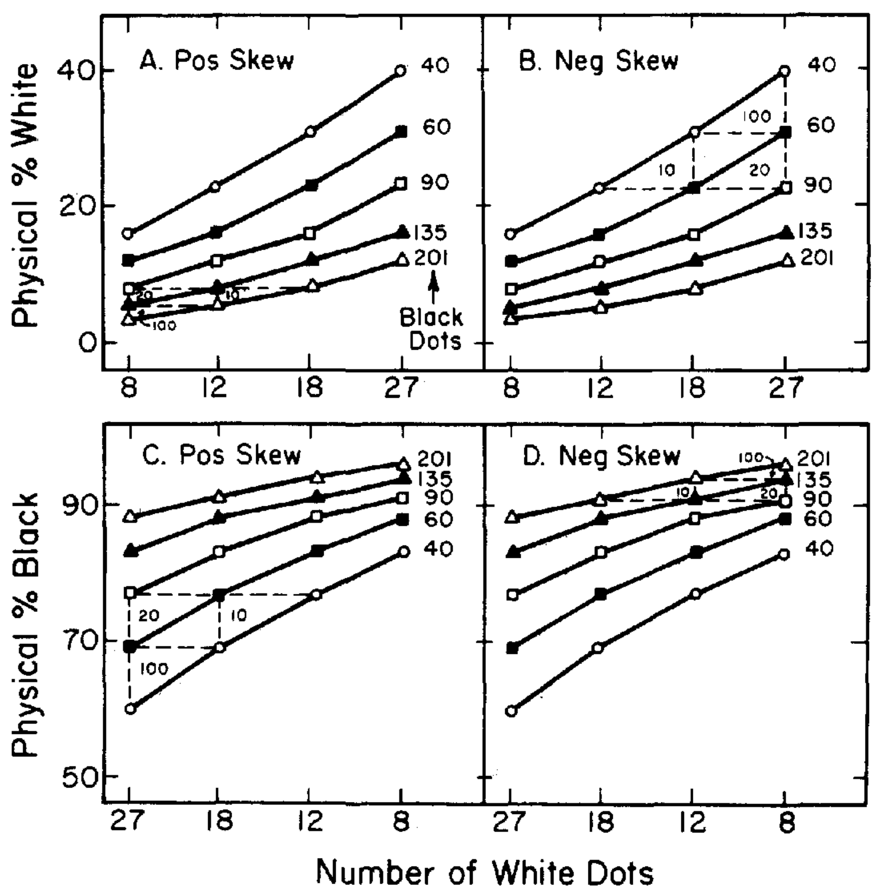

Figure 6. Actual percentages for the four conditions of Experiment 2. (White percentages are shown in the upper panels, and black percentages are shown in the lower panels. The numbers 100, 20, and 10 represent the numbers of contextual stimuli, and their locations depict their composition. Pos = positive; Neg = negative.)

introductory psychology course. There were between 29 and 50 subjects in each of the four conditions. Data from a few additional subjects who did not complete the task were not used in the analyses.

Results

Figure 7 shows mean responses to the "percentage white" tasks in the upper panels and the "percentage black" tasks in the lower panels, plotted as a function of the white dot scale values, with a separate symbol for each number of black dots. Solid lines are the predictions of a theory that will be discussed later. "Percentage white" judgments (upper panels) show divergent interactions between the numbers of white and black dots, as predicted by the relative ratio operation. "Percentage black" judgments (lower panels) show convergent interactions. The majority of individual subjects in each of the four conditions gave responses that matched the divergence or convergence of the means in their condition.

A comparison of panels on the left with those on the right in Figure 7 shows effects of the context. Judged "percentages" for the common stimuli in the positively skewed context are greater than the corresponding "percentages" in the negatively skewed context. Context has a significant main effect: $F(1, 98) = 50.67$ and $F(1, 98) = 17.00$ for white and black "percentages," respectively. For example, the stimulus in Figure 1, which had an actual white percentage of 17%, was called either 21% or 32% depending on the context. In addition, the interaction between black dots and white dots changes, depending on the context. The three-way interaction between white dots, black dots, and context is significant for both "percentage white" and "percentage black" judgments, $F(12, 1176) = 7.54$ and $F(12, 1176) = 2.07$.

Judged proportion versus actual proportion. Figure 8 plots judgments from the common designs (averaged over stimuli with equal physical percentages) against actual percentages for the four conditions. There are three important points to note. First, for both contexts, small "percentages" are overestimated and large "percentages" are underestimated, as in Experiment 1. Individual data were consistent with these trends; in the "percentage white" tasks, 83% of the subjects overestimated small percentages and in the "percentage black" tasks, 92% of the subjects underestimated large percentages.

Second, the relationship between judged "percentages" and actual percentages depends on the context. For "percentage

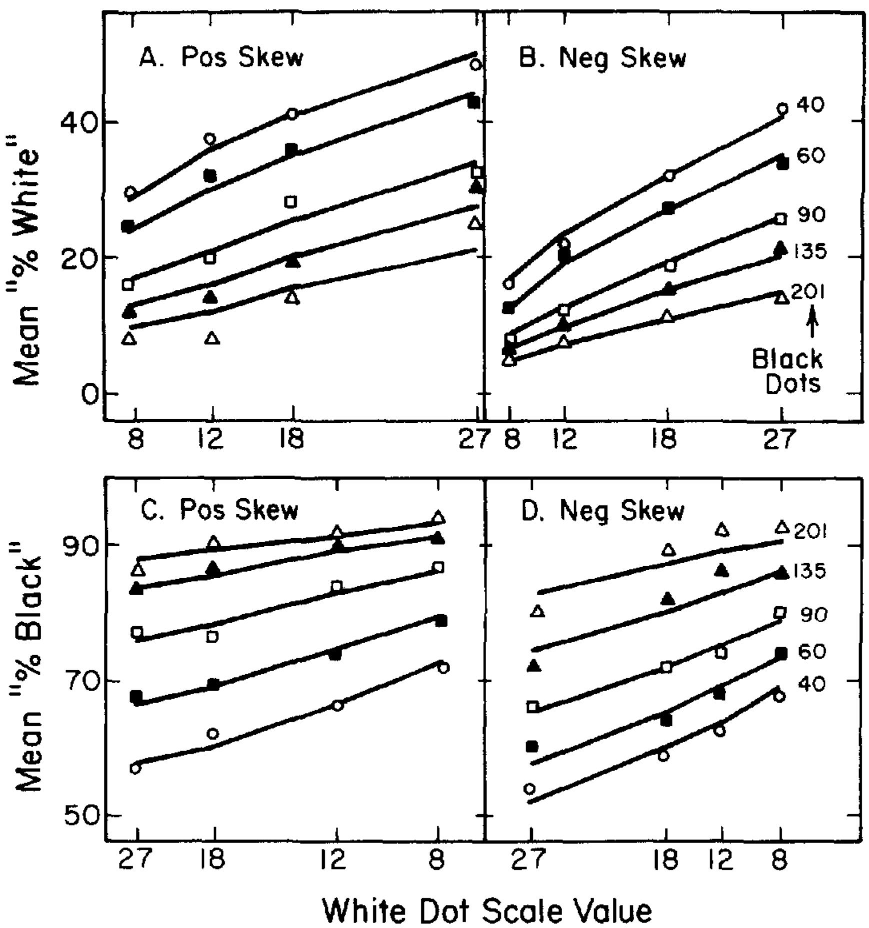

Figure 7. Mean "percentage white" judgments (upper panels) and "percentage black" judgments (lower panels) plotted as in Figure 6. (Lines show predictions of the relative ratio model. Pos = positive; Neg = negative.)

white" judgments (see Figure 8, lower left curves), judgments of small proportions are smaller when there are many proportions near 0.5 (negative skew) than when most stimuli have proportions near zero (positive skew). For "percentage black" judgments (upper right curves), judgments of large proportions are larger when there are many proportions near 0.5 (positive skew) than when there are many contextual proportions near 1.0 (negative skew). These results indicate that judged "percentages" cannot simply be corrected by a single transformation to map them into actual percentages; a theory of the context is required to predict the relationship between judged and actual percentages.

Third, for both "percentage white" and "percentage black" judgments, positively skewed percentages are concave downward in relation to negatively skewed percentages. This shape is predicted by range–frequency theory for these two distributions (Parducci, 1968, 1974; Parducci & Perrett, 1971).

Fit of the relative ratio model. Because of the success of the relative ratio operation for "percentage" judgments in Experiment 1, this model was fit to mean responses for the common stimuli in the four conditions with a special computer program that selected parameters to minimize the sum of the four proportions of errors (Equation 5), as in Experiment 1. White scale values and black scale values were estimated separately for each condition (with the smallest black scale value arbitrarily set to its physical value of 40). The proportions of systematic residual variance in each of the four conditions ranged from .007 to .028. The fit of the model can be assessed in Figures 7 and 8 by comparing predictions of the theory (lines) with data (symbols).

Figure 9 shows the estimated scale values for white dots plotted against physical value. The left and right panels show

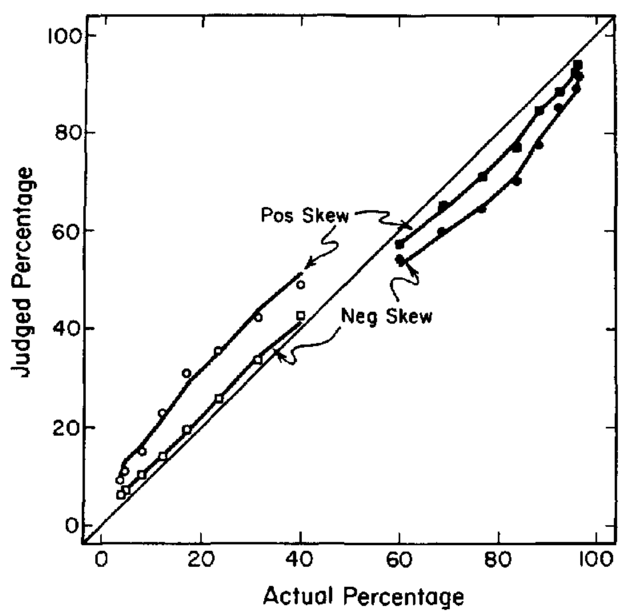

Figure 8. Judged "percentage" against actual percentage for the four conditions. (Open circles are "percentage white" judgments; solid circles are "percentage black" judgments. Lines show predictions of the relative ratio operation, allowing changes in the context to affect the scale values. Pos = positive; Neg = negative.)

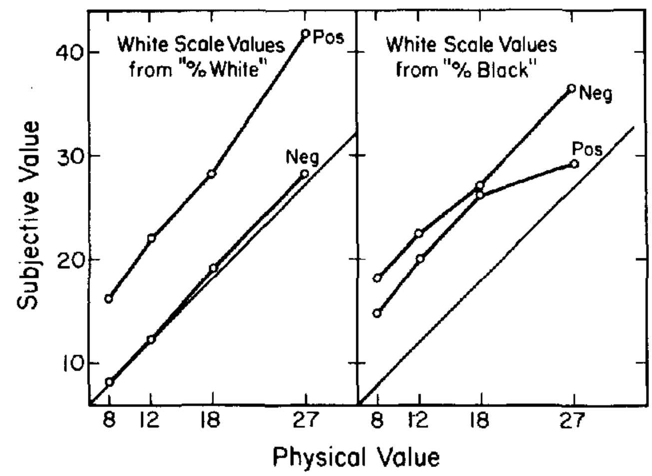

Figure 9. Estimated scale values for white dots estimated from the "percentage white" tasks (left panel) and the "percentage black" tasks (right panel) plotted as a function of physical value. (Pos and Neg refer to positive and negative marginal stimulus distributions, respectively. In the panel on the left, Pos and Neg refer to Panels A and B of Figure 6, respectively. In the panel on the right, Pos and Neg refer to Panels C and D of Figure 6, respectively.)

white dot scale values estimated from the "percentage white" and "percentage black" tasks, respectively, with a separate curve for each context. In Figure 9, context labels refer to the marginal distributions. For white dot scales estimated from "percentage white" tasks, Pos and Neg refer to Panels A and B of Figure 6, respectively. For white dot scales estimated from the "percentage black" tasks, Pos and Neg refer to Panels D and C of Figure 6, respectively.

If there were no effects of either the context or the task, all four curves in Figure 9 would be identical. Instead the curves differ as a function of both task and context. Contextual effects (differences between the positive and negative curves) are more pronounced for white dot scale values when subjects are judging the "percentage" of white dots than when they are judging the "percentage" of black dots. For comparison, white dot scale values in the right panel (estimated from the "percentage black" tasks) show a much smaller difference due to context. Effects of the task were examined by transforming the "percentage black" judgments to "percentage white" judgments (subtracting them from 100%). Examination of task effects showed no systematic main effects. However, there were significant interactions between task and white dots, $F(3, 336) = 12.75$, and between task and black dots, $F(4, 448) = 9.52$.

Figure 10 shows the estimated subjective values for black dots plotted against physical values. The panel on the left shows black scale values from the "percentage black" tasks, with a separate curve for each context. Scale values from the positively skewed distribution of percentages are concave downward in relation to those from the negatively skewed distribution of percentages. Black scale values from the "percentage white" tasks in the right panel are almost identical in the two contexts. Contextual effects on the black scale values

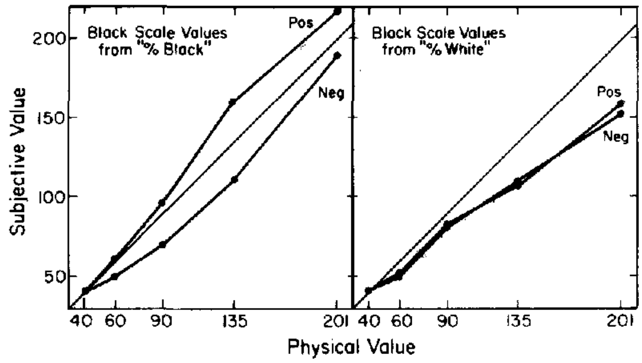

Figure 10. Estimated scale values for black dots estimated from the "percentage black" tasks (left panel) and the "percentage white" tasks (right panel). (Pos and Neg refer to positive and negative marginal stimulus distributions, respectively. In the panel on the left, Pos and Neg refer to Panel C and D of Figure 6, respectively. In the panel on the right, Pos and Neg refer to Panels A and B of Figure 6, respectively.)

are more pronounced in the "percentage black" tasks than in the "percentage white" tasks.

In summary, the data obtained in Experiment 2 appear consistent with the theory that subjects use a relative ratio operation when judging "percentages" and that context effects (variations in the stimulus distribution) influence the scale values. Because both black and white dots are necessary for a percentage judgment, it may seem surprising that contextual effects on the white dot scale values are more pronounced in the "percentage white" tasks than in the "percentage black" tasks. Perhaps when judging "percentage white," subjects separate the white dot context from the black dot context and pay more attention to everything about white dots. This increased attention may result in greater changes in the white dot scales with variations in the context.

Such specific effects of context seem compatible with findings that subjects can make judgments of one type of stimulus that are independent of the distribution of other types of stimuli shown and judged in the same sequence. Parducci, Knobel, and Thomas (1976) instructed a group of subjects to rate the size of each figure in a sequence of squares and circles by judging each square in comparison with the other squares and each circle with the other circles. Variation in the skewness of the distribution of squares did not influence ratings of circles, and vice versa, although the distribution of elements of the same type did affect the judgments.

General Discussion

Results from the present experiments can be summarized as follows:

1. Data from Experiment 1 are consistent with the theory that subjects used the instructed operations with the same scale values in all four tasks. These results can be interpreted as consistent with Birnbaum's (1980, 1982) theory of comparison if it is assumed that subjective zero points are well-defined for the present stimuli.

2. "Percentage" judgments differ from actual percentages in that small percentages are overestimated and large percentages are underestimated. This pattern follows from the relative ratio model if the psychophysical function has the property that the ratio of the subjective values (the smaller element relative to the larger element) is greater than the ratio of their corresponding physical values.

3. "Percentage" judgments depended on the overall stimulus distribution in Experiment 2. Effects of the context appear consistent with Parducci's (1974) range-frequency theory. A model that attributes contextual effects to shifts in the scale values fit the data well. The magnitude of the shift was greater for the color of dots named in the task; scale values for the number of white dots estimated from the "percentage white" judgments changed more as a function of the context than did white dot scale values estimated from the "percentage black" judgments. Similarly, black scale values were more sensitive to the context in the "percentage black" task than in the "percentage white" task.

Subjective Versus Objective Proportion

The pattern of overestimation and underestimation shown in Figures 5 and 8 has been found in a number of studies, including those of Stevens and Galanter (1957) for judgments of proportions, Dale (1959) for gambling behavior, Attneave (1955) for judged frequencies of the occurrence of letters in English text, Begg (1974) for estimates of word frequencies, Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978) for judged frequencies of lethal events (causes of death), and Kellog and Dowdy (1986) for judged frequencies of dots.

The patterns in Figures 5 and 8 are also compatible with a theory of the subjective probability function in risky choice. Karmarkar (1978) noted that certain phenomena in risky decision making could be explained if subjective probability could be described with the equation $s(p) = p^a/[(1 - p)^a + p^a]$, where $s(p)$ is the subjective probability, p is the objective probability, and a is an exponent. When a is less than 1, the relation between subjective and objective probability has a form similar to that shown in Figure 5.

It is interesting to compare the present results with research on calibration. In studies of confidence, people typically answer binary choice questions and provide confidence ratings or estimates of the probability that their answers are correct on a scale from 0.5 to 1.0. For each value of judged confidence, the percentage of correct items is computed. A typical finding is overconfidence: The judged confidence between 0.5 and 1.0 exceeds the actual proportion correct (Fischhoff, Slovic, & Lichtenstein, 1977; Lichtenstein, Fischhoff, & Phillips, 1982).

One difference between calibration studies and the present studies is the method of analysis. When one plots average judged proportion for each level of objective proportion (in the range of 0.5 to 1.0), subjective proportions are typically less than objective proportions. Overconfidence occurs in calibration experiments when one plots average objective proportion for each level of subjective proportion (in the range of 0.5 to 1.0). These two sets of findings can perhaps be reconciled by considering the role of statistical regression.

Regression implies these patterns when the stimulus and response distributions are the same and when the correlation between judged proportion and objective proportion is less than perfect.

Contextual Effects and Between-Subjects Designs

It seems reasonable to assume as a null hypothesis that different types of judgments, including probability judgments, are governed by the same laws. Contextual effects in proportion judgments suggest that probability judgments are described by the same principles that apply to other domains of psychological judgment. Previous research has found that the function relating stimulus to response depends on the range and frequencies of the stimulus distribution and on the range and shape of the response distribution, according to extensions of Parducci's (1968, 1974) range–frequency theory. Although Parducci's theory was originally proposed for category ratings, it appears that contextual effects found in proportion judgments are similar to those found with other judgments, including magnitude estimation, absolute numerical estimation, and other numerical scales, as well as category ratings (Mellers, 1982, 1986; Mellers & Birnbaum, 1982).

In the present experiments with proportion judgments, the stimulus distribution was manipulated but the response distribution was not. The response distribution can be manipulated in magnitude estimation or "ratio" estimation by varying the examples mentioned incidentally in the instructions. For example, Mellers et al. (1984) found that subjects would judge the ratio of 290 grams to 20 grams as either 8 or 32, if the largest example response was any of those values. Hardin and Birnbaum (1990) found that the judged "ratio" of the prestige of a physician to that of a trash collector was either 4 or 64, depending on whether the largest example response was a "ratio" of 4 or 64.

Perhaps the extreme malleability of "ratio" judgments can be explained by the fact that subjects are really computing differences when instructed to judge such "ratios" for those continua. On the other hand, proportion judgments might not be as easy to manipulate because subjects are thought to be using the same operation as instructed when making their judgments. The response distribution in proportion judgment might be influenced by examples (e.g., "If you think 1 person in 6 has diabetes, say 1/6" vs. "If you think 1 person in 10,000 has diabetes, say 1/10,000"). Nevertheless, it seems plausible that if the stimulus distribution affects judged proportions, the response distribution could also have such effects.

The important question is whether judgments of probability can be taken at face value. The fact that the same physical proportion can receive different judgments in different contexts suggests that the results in Figures 5 and 8 have implications for theories of intuitive probability and also for the interpretation of studies that involve judgments of probabilities.

To understand probability judgments, it is necessary to distinguish among three concepts: subjective, objective, and judged probability. Some investigators have concluded that because probability is on an absolute scale, judgments of probability are also absolute and are identical to subjective probabilities. Those who treat judged probability as identical to subjective probability might be tempted to conclude from between-subjects comparisons that if an event receives a higher probability judgment, then it also has a higher subjective probability. However, Figures 5 and 8 show that a theory of the context is necessary to compare probability judgments between different groups of subjects. For example, by comparing judgments between groups, Figure 8 shows that the physical proportion of .17 is judged as higher than the physical proportion of .23 (30.5 vs. 25.3, the fifth open circle and the sixth open square from the left). In contrast, a within-subject comparison shows that both groups of subjects gave higher judgments to higher physical proportions. Because between-subjects comparisons (Figure 8) can lead to such obvious contradictions, one should be extremely cautious when drawing inferences from between-subjects comparisons of judgments.

Birnbaum and Mellers (1983) found that research on the "base-rate fallacy" leads to different conclusions depending on whether the experiments used within-subject or between-subjects designs. The interpretation of base rate "neglect" is based on the finding that in between-subjects comparisons, the effect of base rate is too small (Kahneman et al., 1982). In within-subject comparisons, however, subjects use the base rate, and the evidence for a "fallacy" disappears (Birnbaum & Mellers, 1983). Birnbaum (1982) noted that the judged fault of a rape victim also differs in within-subject and between-subjects designs, and he gave a range–frequency analysis of why between-subjects comparisons lead to paradoxical conclusions: Between-subjects comparisons confound the stimulus and the context by allowing the stimulus to evoke its own context. Nihm's (1984) satire on the claim that subjects lack self-insight also shows how unusual conclusions could be reached from between-subjects comparisons. If we do not wish to argue that larger physical proportions are judged smaller, the present results provide another argument for preferring within-subject comparisons.

If there were a single function relating judged probability to objective probability, one could simply apply the inverse function to convert judgments of probability into "real" probabilities. The present data demonstrate that no single function would suffice, because such corrections depend on the context. Furthermore, it has been shown that judgments of probability do not obey the algebra of probability, even if one allows for a monotonic transformation of the judgments (Birnbaum & Mellers, 1983). Therefore, even subjective probabilities do not obey the algebra that would justify the construction of an absolute scale.

In principle, it is possible to determine whether contextual effects can be attributed to the response scale or to the perception of the stimulus (Mellers & Birnbaum, 1982); however, the present data do not provide an unambiguous determination of the locus of the effect. Although the data could be well fit by assuming that contextual effects operate on the scale values, in other judgment tasks contextual effects have been attributed to the response stage (Mellers & Birnbaum, 1983).

An analogous (but perhaps more philosophical) question is to ask if the contextual effects are truly psychological or merely semantic. This ill-defined question leads one to speculate about related experiments. For example, suppose the subject's task was to make bids for the opportunity to play gambles. Suppose a ball was drawn from an urn represented by Figure 1, and that if the ball was white, the subject would win $100 and if the ball was black, the subject would receive nothing. Would subjects in the positively skewed condition offer more to play such a gamble than subjects in the negatively skewed condition? If the judgments in Figure 8 are indicative of true subjective probabilities, then increasing the subjective probability of winning should increase the value of the gambles based on those probabilities. Recent research shows that the judged value of a gamble depends on the distribution of gambles presented for judgment (Mellers, Ordonez, & Birnbaum, 1989).

Conclusions

The present experiments are consistent with the theory that subjects use the instructed operation when judging "ratios," "differences," and "percentages." Despite the use of the correct operation, subjects overestimate small percentages and underestimate large percentages. Percentage judgments show contextual effects that are due to changes in the stimulus distribution; the same physical proportion receives different judgments in different contexts. These experiments suggest that proportion judgments are governed by principles similar to those found with other numerical responses. Thus, proportion judgments and, by analogy, probability judgments should not be taken at face value because they depend on both the stimulus information and the surrounding context.

References

Attneave, F. (1955). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology, 46,* 81–86.

Beckwith, M., & Restle, F. (1966). Process of enumeration. *Psychological Review, 73,* 437–444.

Begg, I. (1974). Estimation of word frequency in continuous and discrete tasks. *Journal of Experimental Psychology, 102,* 1046–1052.

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

Birnbaum, M. H. (1978). Differences and ratios in psychological measurement. In N. J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 33–74). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1980). Comparison of two theories of "ratio" and "difference" judgments. *Journal of Experimental Psychology: General, 109,* 304–319.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401–485). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1983). Base rate in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology, 96,* 85–94.

Birnbaum, M. H., Anderson, C., & Hynan, L. G. (1989). Two operations for "ratios" and "differences" of distances on the mental map. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 785–796.

Birnbaum, M. H., & Elmasian, R. (1977). Loudness "ratios" and "differences" involve the same psychophysical operation. *Perception & Psychophysics, 22,* 383–391.

Birnbaum, M. H., & Mellers, B. A. (1978). Measurement and the mental map. *Perception & Psychophysics, 23,* 403–408.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792–804.

Birnbaum, M. H., & Veit, C. T. (1974). Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception & Psychophysics, 16,* 276–282.

Chandler, J. P. (1969). STEPIT: Finds local minima of a smooth function of several parameters. *Behavioral Science, 14,* 81–82.

Dale, H. C. (1959). A priori probabilities in gambling. *Nature, 183,* 842–843.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmutz (Ed.), *Formal representations of human judgment* (pp. 17–52). New York: Wiley.

Elmasian, R., & Birnbaum, M. (1984). A harmonious note on pitch: Scales of pitch derived from the subtractive model of comparison agree with the musical scale. *Perception & Psychophysics, 36,* 531–537.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564.

Hardin, C., & Birnbaum, M. H. (1990). Malleability of "ratio" judgments of occupational prestige. *American Journal of Psychology, 103,* 1–20.

Hasher, L., & Zacks, R. (1984). Automatic processing of fundamental information. *American Psychologist, 39,* 1372–1388.

Jensen, E. M., Reese, E. P., & Reese, T. W. (1950). The subitizing and counting of visually presented fields of dots. *Journal of Psychology, 30,* 362–392.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature, 3,* 281–282.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance, 21,* 61–72.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *American Journal of Psychology, 62,* 498–525.

Kellog, R. T., & Dowdy, J. C. (1986). Automatic learning of the frequencies of occurrence of stimulus features. *American Journal of Psychology, 99,* 111–126.

Klahr, D. (1973). Quantification processes. In W. G. Chase (Ed.), *Visual information processing* (pp. 3–34). Orlando, FL: Academic Press.

Klahr, D., & Wallace, J. G. (1973). The role of quantification operators in the development of conservation of quantity. *Cognitive Psychology, 4,* 301–327.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, F., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–578.

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

Mellers, B. A. (1986). "Fair" allocations of salaries and taxes. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 80–91.

Mellers, B. A., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 582–601.

Mellers, B. A., & Birnbaum, M. H. (1983). Contextual effects in social judgment. *Journal of Experimental Social Psychology, 19,* 157–171.

Mellers, B. A., Davis, D., & Birnbaum, M. H. (1984). Weight of evidence supports one operation for "ratios" and "differences" of heaviness. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 216–230.

Mellers, B. A., Ordonez, L., & Birnbaum, M. H. (1989). *Context and response mode effects in risky decision making.* Unpublished manuscript.

Nihm, S. D. (1984). Self reports on mental processes: Comment on Birnbaum and Stegner. *Bulletin of the Psychonomic Society, 22,* 426–427.

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

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 127–141). Orlando, FL: Academic Press.

Parducci, A., Knobel, S., & Thomas, C. (1976). Independent contexts for category ratings: A range-frequency analysis. *Perception & Psychophysics, 20,* 360–366.

Parducci, A., & Perrett, L. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology Monograph, 819,* 427–452.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29–46.

Philip, B. R. (1947). Generalization and central tendency in the discrimination of a series of stimuli. *Canadian Journal of Psychology, 1,* 196–204.

Schum, D. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimony. *Organizational Behavior and Human Performance, 27,* 153–196.

Shanteau, J. (1974). Component processes in risky decision making. *Journal of Experimental Psychology, 103,* 680–691.

Shuford, E. H. (1961). Percentage estimation of proportion as a function of element type, exposure time, and risk. *Journal of Experimental Psychology, 61,* 430–436.

Slovic, P., Lichtenstein, S., & Fischhoff, B. (1988). Decision making. In R. C. Atkinson, R. J. Hernstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (2nd ed., pp. 673–738). New York: Wiley.

Stevens, S. S., & Galanter, E. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology, 54,* 377–411.

Treisman, A. (1988). Feature analysis in early vision: Evidence for search asymmetries. *Psychological Review, 95,* 15–48.

Veit, C. T. (1978). Ratio and subtractive processes in psychophysical judgment. *Journal of Experimental Psychology: General, 107,* 81–107.

Wallsten, T. S., Budescu, D. V., Rappoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meaning of probability terms. *Journal of Experimental Psychology: General, 115,* 348–365.

Zadeh, L. A. (1975a). The concept of a linguistic variable and its application to approximate reasoning: Part 1. *Information Science, 8,* 199–249.

Zadeh, L. A. (1975b). The concept of a linguistic variable and its application to approximate reasoning: Part 2. *Information Science, 8,* 301–357.

Zadeh, L. A. (1975c). The concept of a linguistic variable and its application to approximate reasoning: Part 3. *Information Science, 9,* 43–98.

Zimmer, A. C. (1983). Verbal vs. numerical processing of subjective probabilities. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 159–182). Amsterdam: North-Holland Publishers.

Received March 21, 1989
Revision received July 18, 1989
Accepted January 2, 1990 ■