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

2

Differences and Ratios in Psychological Measurement

Michael H. Birnbaum
University of Illinois at Urbana-Champaign

The purpose of this chapter is to present new approaches to an old problem in psychophysics: the apparent contradiction between so-called "ratio" and "interval" techniques for scaling. The quote marks are used to remind the reader that the numbers obtained when subjects are instructed to judge "ratios" or "differences" need not obey the mathematical properties of the ratio or subtractive models.

The chapter begins with a brief discussion of "direct" scaling, the so-called new psychophysics that was to replace the scaling methods of Fechner and Thurstone. One major problem with this approach is that the empirical contradiction in scales cannot, in principle, be resolved within the unifactor framework typically used in "direct" scaling, which does not allow tests of the theories of measurement.

Newer approaches of psychological measurement that use factorial stimulus designs and algebraic models to assess the data are discussed. It is shown that the additional criterion of scale convergence, the premise that stimulus scales are independent of instructions, can add the extra constraint needed to resolve certain indeterminacies in the measurement approach.

Results of recent research using factorial designs with "ratio" and "difference" tasks illustrate the evidence that for a variety of perceptual continua, instructions to judge "ratios" or "intervals" lead to the same ordering of stimulus pairs, consistent with the interpretation that there is but one comparison operation—which could be either a ratio or difference—for both tasks.

A framework is presented in which ratio and subtractive theories make different ordinal predictions for more complex judgment tasks in which subjects make quantitative comparisons of stimulus pairs. Two experiments are then reviewed in which "ratios" and "differences" of stimulus intervals satisfy predictions of both ratio and subtractive models, yielding a ratio scale of intervals. This scale of intervals is then used to resolve the ratio versus subtractive interpretations for simple "ratios" and "differences" of two stimuli. Data from the two experiments support the contention that the basic operation by which subjects compare two stimuli is best represented by subtraction.

"DIRECT" SCALING

One proposal to obtain a scale of psychological magnitude was to ask subjects to report numbers that "directly" represent the strengths of sensations. The term "direct" was used to emphasize the distinction between this technique and the Fechner-Thurstone approach of "indirectly" inferring psychological differences from measures of discriminability (Stevens, 1957).

An outline of "direct" scaling is shown in Fig. 1. In the outline, Φ_i is the physical measurement of stimulus level i, s_i is the corresponding sensation, and R_i is the overt numerical response. The function relating sensations to physical values is termed the *psychophysical function*, $s_i = H(\Phi_i)$. The function relating responses to subjective values is called the *judgment function*, $R_i = J(s_i)$. A plot of responses against the physical values represents the composition $R_i = J[H(\Phi_i)]$.

Examples of "Direct" Scaling

The bottom of Fig. 2 shows seven squares containing dot patterns that were used as stimuli in an experiment to illustrate typical results obtained with "direct" scaling methods. Subjectively, how "dark" are the dot patterns? Two "direct" methods have been used in attempts to answer this question. The first is to ask subjects to produce numbers that represent subjective intervals using the method of *category rating.* A second procedure is to ask subjects to report numbers that are "directly proportional to their sensations," a technique called *magnitude estimation.* The numbers obtained by these two "direct" methods constitute two operational definitions of "sensation."

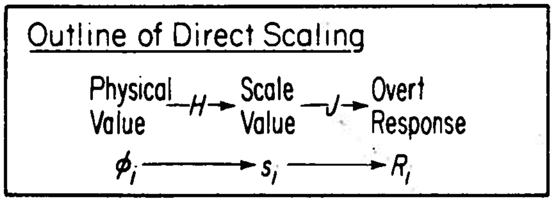

FIG. 1 Outline of "direct" scaling. Physical values of the stimuli (Φ_i) are related to psychological scale values (s_i) by the psychophysical function, $s_i = H(\Phi_i)$. Overt responses, R_i, are related to subjective values by the judgment function, $R_i = J(s_i)$, assumed to be strictly monotonic. Since the data observed in a typical unifactor "direct" scaling study are the confounded composition, $R_i = J(H(\Phi_i))$, it is not possible to separate theories of subjective value, comparison processes, or judgmental processes in this framework.

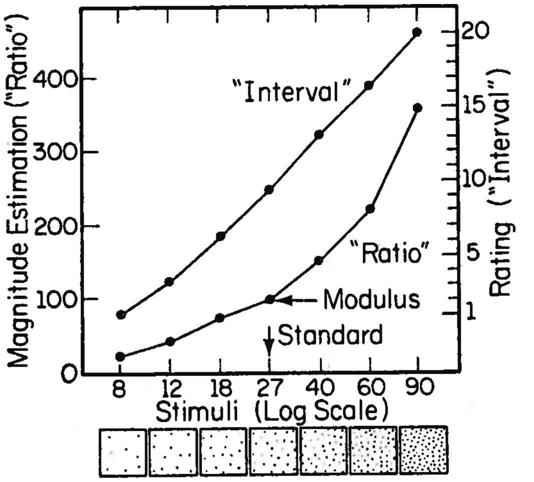

FIG. 2 "Direct" scalings using magnitude estimations and category ratings yield different "scales" of sensation. Upper curve shows mean ratings (1-20) of darkness of dot patterns as a function of the number of dots, spaced logarithimically on the abscissa. Lower curve shows mean magnitude estimations, when the 27-dot pattern (standard "divisor") was designated "100" (modulus).

One group of subjects was instructed to make magnitude estimations of "ratios." They were asked to call the fourth pattern (the standard) "100" and to assign numbers to the other patterns so that the ratio of the two numerical responses would equal the ratio of the subjective darknesses of the two stimuli. They were instructed, if it seemed half as dark, to say "50," if it seemed twice as dark, to say "200," etc.

A second group was asked to make ratings that would represent "intervals." They were told to call the lightest pattern "1," the darkest "20," and to judge each square so that the differences in response would be proportional to subjective intervals of darkness.

Assuming that the subjects follow these instructions, we theorize that the magnitude estimation of "ratio" should be given by the equation:

$$\mathrm{ME}_i = 100\left(\frac{s_i}{s_4}\right) \tag{1}$$

where s_4 is the subjective darkness of stimulus 4 (the standard), s_i is the sensation for stimulus i, and ME_i is the overt magnitude estimation.

Similarly, the category judgments should be given by the equation:

$$\mathrm{CJ}_i = 19\left(\frac{s_i - s_1}{s_7 - s_1}\right) + 1, \tag{2}$$

where CJ_i is the category judgment, and s_1 and s_7 are the two extreme stimuli to be judged "1" and "20," respectively.

Equations (1) and (2) represent theories of the judgment function, J, although they involve deeper assumptions about comparison processes which are discussed later. Equation (1) asserts that magnitude estimation responses are directly proportional to subjective values; Eq. (2) asserts that category ratings are linearly related to subjective values. If the subjective values (s_i in both equations) are the same, solving Eq. (1) for s_i and substituting in Eq. (2) shows that category ratings should be linearly related to magnitude estimations:

$$\mathrm{CJ}_i = \frac{19}{s_7 - s_1}\left[\frac{s_4}{100}(\mathrm{ME}_i) - s_1\right] + 1 \tag{3}$$

Equation (3) expresses the idea of the convergence of two operational definitions of sensation. If two procedures for defining the same construct were to agree, there would be no evidence for concern. However, when two procedures do *not* agree, an explanation is required.

An Empirical Contradiction

Figure 2 plots both magnitude estimations ("ratios") and category ratings ("Intervals") as a function of the number of dots in the squares, spaced in equal log steps on the abscissa. It should be apparent that the two procedures yield scales that are not linearly related. These results are typical of the results obtained in a large number of experiments for a variety of psychophysical and social judgment dimensions (Stevens & Galanter, 1957; Stevens, 1966). Instead of a linear function, magnitude estimations are often approximately exponentially related to category ratings.

This violation of converging operations, though it caused some consternation, had good effects for the study of psychological scaling. It caused psychologists to argue about methods, theories, and data, and it caused them to doubt the meaningfulness of the entire enterprise of psychological scaling based on operational definitions (Treisman, 1964; Savage, 1966). Theories were proposed to account for the glaring discrepancy between the two methods. Unfortunately, the theories were untestable in the traditional framework of "direct" scaling.

Theories of the Discrepancy

Three general theories were proposed to account for the finding that the "interval" scaling techniques gave results that were nonlinearly related to the "ratio" techniques: (1) the judgment function, *J*, depends on the response procedure and is nonlinear for at least one of the methods; (2) there is some bias in the comparison process, *C*, so that subjects cannot compute both ratios and differences properly; and (3) the subjective values of the stimuli, *s*, change value, depending on the task.

The first theory is that at least one of the procedures contains a nonlinear judgment bias. Thus this theory rejects the assumptions [of Eqs. (1) and (2)] that *J* is a linear function in the case of ratings and *J* is a similarity transformation in the case of magnitude estimation. Just because the subjects have been *instructed* to report a number that is directly proportional to their sensation does not mean that they can use numbers in this way. Attneave (1962) proposed that subjective magnitude of numerals could be a nonlinear power function of objective numerical magnitude, suggesting that the subject selects a magnitude estimation number whose subjective value is equal to the subjective value of the stimulus. Rule, Curtis, and their associates have pursued Attneave's suggestion that the inverse of J is the psychophysical function for numerals (see Rule & Curtis, 1973). Treisman (1964) and Ekman and Sjoberg (1965) have discussed the possibility of a logarithmic psychophysical function for numerals, which would produce an exponential function for J. Poulton (1968) attributed the nonlinearity of J to context effects in the experiments, since the results of "direct" scaling studies using magnitude estimation depended on the stimulus range, frequency distribution, value of the standard, and a variety of other experimental details.

The second type of theory contends that at least one computation is erroneous. Stevens (1971) argued, for example, that subjects can estimate ratios but cannot compute subjective intervals. Rather than argue for a computation error in one of the operations, Torgerson (1961) advanced the conjecture that subjects do not distinguish between "ratios" and "differences," perceiving instead only one quantitative comparison between a pair of stimuli. Subjects (and the experimenters) are willing to call this single relationship either a "ratio" or a "difference."

A third possibility is that the sensation depends on the task. This position contends that we should take the responses in Fig. 2 at face value and conclude that there are at least two kinds of sensations that are nonlinearly related. Marks (1974) has argued that there are two different scales of sensory magnitude, one for "intervals" and one for "ratios," related by the square root function.

Problems with "Direct" Scaling

The major problem with the traditional method of "direct" scaling is that the experimental designs do not provide sufficient ordinal constraints to test the theories of stimulus comparison (Krantz & Tversky, 1971; Birnbaum & Veit, 1974a; Anderson, 1974; Veit, 1974; Shepard, 1976). "Ratios" and "differences" may or may not be consistent with the metric or ordinal predictions of ratio and subtractive models.[1] However, with a unifactor design, it would not be possible to reject a ratio or difference model.

[1] Quotation marks are used throughout to indicate "ratio" and "difference" tasks or judgments. Quotation marks are not used for actual (computed) ratios and differences or for models and theoretical statements.

With unifactor designs, used in "direct" scaling research (e.g., Stevens & Galanter, 1957; Torgerson, 1960), "ratios" and "intervals" are necessarily monotonically related. Since the standard (divisor) is fixed, $\text{ME}_j = J_{\text{R}}\ (s_j/c)$, where c is the standard, and J_{R} is the judgment function for magnitude estimation. For "interval" judgments,

$$\text{CJ}_j = J_{\text{D}}\left(\frac{s_j - s_1}{s_7 - s_1}\right),$$

Where s_1 is the smallest stimulus, s_7 is the largest, and J_{D} is the judgment function for ratings. Hence, since ME_j and CJ_j are both monotonic functions of s_j, ME_j is monotonically related to CJ_j whether the subject computes a difference or a ratio. Therefore, unifactor designs do not permit ordinal tests of theories, such as Torgerson's (1961), that there is only one comparison operation. However, ratios and differences are *not* monotonically related in general (e.g., $2/1 > 7/5$, but $2\text{-}1 < 7\text{-}5$). With factorial designs it becomes possible to test theories of comparison processes.

Poulton (1968) has noted that the results of direct scaling studies depend on contextual details of experimental procedure. If the responses are taken at face value, it means that the scale of sensation depends on the stimulus context. But with the direct scaling approach, there is no way to test the alternative theory that the context affects only the judgment, not the sensation.

Basically, since observed data are of the form $R = J\ [H\ (\Phi)]$, it is clear that for any reasonable theory of H, it is possible to find a function J such that the composition matches the data in a unifactor design. There is no way to test whether the computation operation can be represented by a difference or a ratio or whether J is linear. There is no way to test whether context affects H or J. The use of factorial designs in the framework of algebraic measurement theories comes close to solving these problems. Certain difficulties still remain but become resolvable with additional constraints. The following section explains how the additional ordinal constraints produced by factorial designs permit tests of algebraic theories of judgment, which allow for distinction between the stimulus scale, $s = H\ (\Phi)$, and the response transformation, $R = J\ (\Psi)$.

PSYCHOLOGICAL MEASUREMENT APPROACH

Figure 3 shows an outline of psychological measurement that facilitates discussion of theories of stimulus comparison. In the outline the comparison operation is represented by the function, $\Psi_{ij} = C\ (s_i, s_j)$, where Ψ_{ij} is the subjective impression of a difference or ratio, s_i and s_j are the subjective scale values of the stimuli, and C is a model of the comparison (integration) process that describes how two stimuli combine to produce the impression of the pair relationship. The judgment function, $R_{ij} = J\ (\Psi_{ij})$, represents the transformation from impression, Ψ, to overt numerical response.

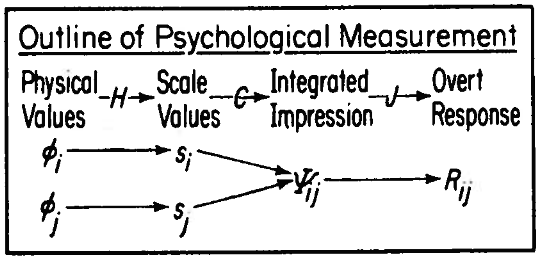

FIG. 3 Outline of psychological measurement. Subjective scale values of the stimuli are combined by the comparison function, $\Psi_{ij} = C\ (s_i, s_j)$, and transformed to an overt response by the strictly monotonic judgment function, $R_{ij} = J\ (\Psi_{ij})$. In this framework, using factorial stimulus designs, the assumption of a linear J permits a metric test of C. Alternatively, the assumption of a theory of C permits estimation of J and the scale values, s.

In this framework it is possible in principle to test (i.e., reject) models of the comparison process. Conjoint measurement analysis (Krantz et al., 1971; Krantz & Tversky, 1971) attempts to specify the ordinal relationships that ideal data must satisfy in order to be consistent with particular theories of C. The approach of functional measurement (Anderson, 1970, 1974) has been to attempt to specify theories that reproduce the metric information in the numerical data that are obtained.

Illustrative Factorial Experiments

Figure 4 represents a factorial stimulus design that is used to convey several ideas of psychological measurement. The reader is invited to make two copies of Fig. 4 and to particpate as a psychophysical observer. For one unfamiliar with this area of research, in order to gain a better grasp of the results that follow, it will be helpful to carry out the analyses described below on a set of data. Although these dot experiments are intended for illustrative purposes, they are convenient and reliable demonstrations of results obtained in more formal experimental settings with other psychophysical continua.[2]

Two experiments illustrate important points of the present chapter. For the "ratio" task, judge the ratio of the darkness of each column square to the darkness of every row square. The estimations should be written in the appropriate matrix locations. Judgments should be consistent with the following scheme:

12.5	=	column is 1/8 as dark as row
25	=	column is 1/4 as dark
50	=	column is 1/2 as dark
100	=	column equals row
200	=	column is 2 times as dark
400	=	column is 4 times as dark
800	=	column is 8 times as dark

[2] In factorial $B \times A$ designs, B refers to rows, indexed by i; A refers to columns, indexed by j. The tasks always specify $A - B$ or A/B, and the data are always plotted against factor A, with a separate curve for each level of B.

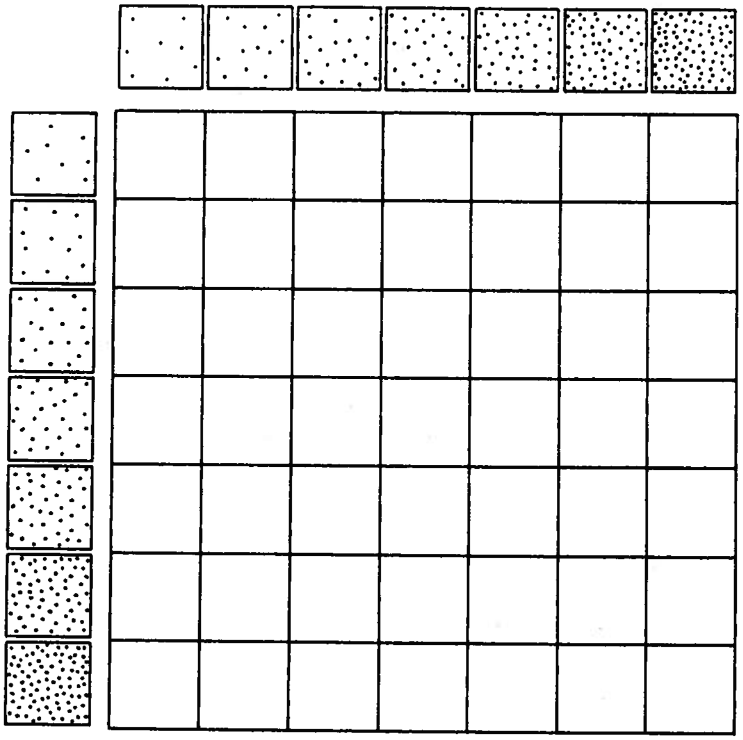

FIG. 4 Representation of factorial experiment. The reader is encouraged to make two copies of this figure and replicate the experiments. For the “ratio” task, judge the ratio of the darkness of the column square to that of the row using a modulus of 100. For the “difference” task, judge the algebraic difference, column minus row, calling the largest difference ‘100,” a zero difference “0,” and using negative numbers when the row stimulus is darker.

Any numbers consistent with this scheme may be used to represent the ratios of subjective, psychological darkness of the dot patterns.

For the “difference” task, the instructions are to judge the difference between the darkness of the column stimulus and the row stimulus. The greatest difference (column-row) is to be called “100.” When the column equals the row darkness, the difference would be “0.” Negative numbers would be used to express negative differences (i.e., when the row stimulus seems darker than the column stimulus).

The two experiments produce two matrices of numbers that correspond to differential instructions to judge “ratios” and “differences.” The following subsections show how the measurement approach permits one to test corresponding ratio and subtractive models and to estimate scale values from the data without using the physical measurements of the stimuli.

Ratio and Subtractive Models

Ratio model. The *ratio* model can be written:

$$R_{ij} = J_{\mathrm{R}} (s_j/s_i), \tag{4}$$

where R_{ij} is the numerical estimation of the "ratio" of the jth column stimulus to the ith row stimulus (divisor), having subjective scale values s_j and s_i, respectively, and J_{R} is the judgmental transformation that relates overt numerical estimations to subjective ratios.

Subtractive model. The *subtractive* model can be written:

$$D_{ij} = J_{\mathrm{D}}\,(s_j - s_i), \tag{5}$$

where D_{ij} is the rated difference between the stimulus of column j and the stimulus of row i, s_j and s_i are the scale values, and J_{D} represents the judgment function that relates overt ratings of differences to subjective intervals.

Metric Implications of the Models

It is useful to initially examine the metric (i.e., numerical) implications of the ratio and subtractive models under the special assumption that the judgment functions, J_{R} and J_{D}, are linear. More complex cases that do not restrict the form of the judgmental transformations are discussed later.

Figure 5 shows computed ratios and differences for a 7 X 7 design, as in Fig. 4. The scale values for the seven rows and columns are assumed to be successive integers from 1 to 7 (i.e., let $s_i = i$, and $s_j = j$), and the judgment functions are assumed to be identity functions. Hence $R_{ij} = j/i$, and $D_{ij} = j - i$. Although for simplicity the scale values and judgment functions have been assumed to be known in this example, the measurement approach allows them to be estimated from the data.

Ratio model. The left-hand panel of Fig. 5 plots ratios, R_{ij}, as a function of the column scale value ($s_j = j$), with a separate curve for each row stimulus. The highest curve represents the first row of the matrix, $R_{1j} = j/1$. The lowest curve represents the last row of the matrix, $R_{7j} = j/7$. Each curve is a linear function of the scale value of the column stimulus, with the same (zero) intercept. The slopes are inversely proportional to the scale values of the row stimuli. This sort of diverging fan of straight lines that intersect at a common point is termed a *bilinear fan,* since the interaction in the matrix is located entirely in the bilinear component. Each entry in the matrix could be produced from the equation,

$$R_{ij} = R_{i.}\, R_{.j} / R_{..}\,, \tag{6}$$

where $R_{i.}$ and $R_{.j}$ are the row and column totals respectively, and $R_{..}$ is the grand total of the matrix. [This equation is analogous to the method for computing predictions under the hypothesis of independence (multiplicative probabilities) for a chi-square table.]

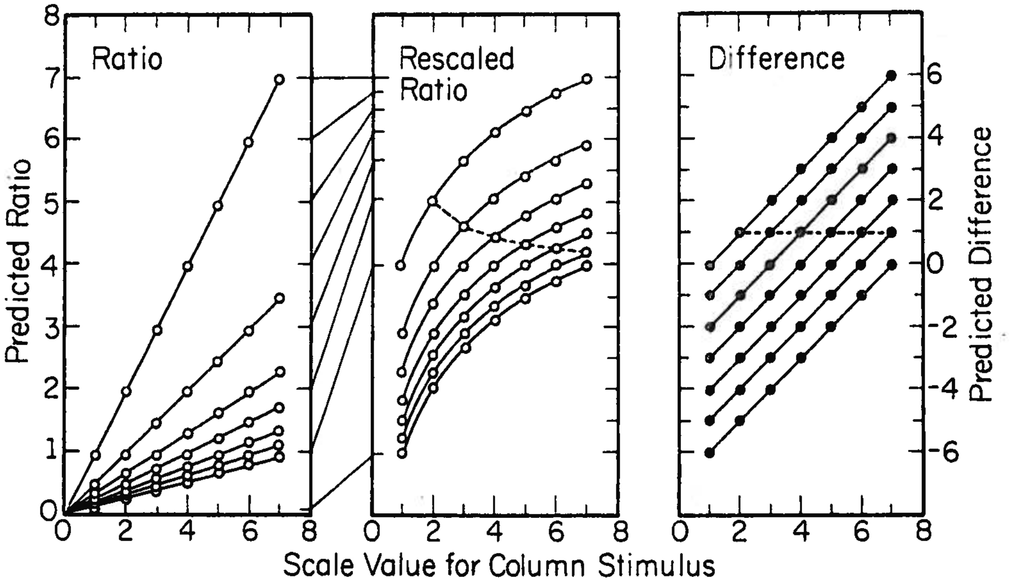

FIG. 5 Predicted ratios and differences assuming two operations on one scale of sensation. The computations have been made using successive integers from 1 to 7 as scale values. *Left:* Plots ratios A/B as a function of the column value (dividend–A) with a separate curve for each row stimulus (divisor–B). *Center:* Shows the ratios rescaled to parallelism by a logarithmic transformation (lines connecting panels). *Right:* Plot differences ($A - B$) as a function of column scale value (minuend), with a separate curve for each row stimulus (subtrahend). Note that ratios and differences are *not* monotonically related. Dashed lines connect differences of 1 and corresponding log ratios.

Equation (6) provides a means for estimating scale values when they are unknown. If there are r rows and k columns, and if $R_{ij} = s_j/s_i$, then

$$R_{.j} = s_j\left(\sum_{i=1}^{r} \frac{1}{s_i}\right)$$

Hence the marginal sum, marginal mean (Anderson, 1970), or $R_{.j}/\sqrt{R_{..}}$ are all proportional to the scale value, s_j. [The last expression may be recognized as the equation for the first centroid factor of a correlation matrix; it is also presented by Ekman (1958).]

If J_{R} were a linear function, $R_{ij} = as_j/s_i + c$, the value of the additive constant, c, could be determined from the projection of the point of intersection onto the ordinate. If J_{R} were of the form, $R_{ij} = a\,(s_j/s_i)^{\,b} + c$, the data would still plot as a bilinear fan, since one could define new scale values, $s' = s^b$, against which the curves would be linear. Scales derived from a ratio model are unique to a power transformation, $s' = as^b$, where a and b are arbitrary (Krantz et al., 1971). [The letters a, b, and c are used throughout this chapter for arbitrary constants, with no carryover from equation to equation.]

Some confusion was created by the assertion that marginal means are an "interval scale" of the stimuli in the ratio model. The term "interval scale" has two meanings. Assuming a linear judgment function, the marginal means are linearly related to (and hence an "interval scale" of) the scale values for the ratio model (Anderson, 1970). The concept of "interval scale" has another meaning in terms of ordinal uniqueness. The ratio model does *not* define an interval scale of the stimuli in the ordinal uniqueness sense, since any power transformation of the scale values would yield scales that would reproduce the rank order of the data. That is, substituting $s' = as^b$ for s in Eq. (4), $R_{ij} = J_{\mathrm{R}}\ (s_j^b/s_i^b) = J_{\mathrm{R}}\ [(s_j/s_i)^b]$. Since the composition of J_{R} and a power function is a monotonic function, the rank order of R_{ij} is preserved by any power transformation of the scale values.

In summary, the ratio model predicts a bilinear fan of curves that intersect at a common point. This prediction depends on the judgment function being of the form: $R_{ij} = a\ (s_j/s_i)^b + c$. When the pattern of bilinearity is observed, it is possible to estimate scale values for the ratio model from marginal sums or means. The scales thus defined are unique to a power transformation.

Subtractive model. The right-hand panel of Fig. 5 shows the computed differences, $D_{ij} = j - i$, plotted in the same fashion as the ratios in the left-hand panel. Again, the curves are linearly related to the column scale value. In the case of the subtractive model, however, the curves are parallel. If $D_{ij} = a\ (s_j - s_i) + c$, then $D_{ij} - D_{kj} = a\ (s_k - s_i)$ for all j. Hence the difference between any two rows, say i and k, is independent of the column j.

The parallelism implied by the subtractive model is equivalent to the condition of no row X column interaction in the analysis of variance. Under these conditions, each entry in the matrix can be reproduced by the equation,

$$D_{ij} = \bar{D}_{i\cdot} + \bar{D}_{\cdot j} - \bar{D}_{\cdot\cdot}, \tag{7}$$

where $\bar{D}_{i\cdot}$ *and* $\bar{D}_{\cdot j}$ are the marginal means, and $\bar{D}_{\cdot\cdot}$ is the grand mean in the matrix.

Equation (7) provides a means of estimating scale values, since $\bar{D}_{\cdot j} = a\ [s_j - \Sigma\ (s_i/r)] + c$. Therefore, when the data fit the model, the curves are parallel and the marginal means are linearly related to the subtractive model scale values. If J_{D} were a linear function, $D_{ij} = a\ (s_j - s_i) + c$, the curves would remain parallel. The scale values for the subtractive model are unique to an interval scale, since any linear transformation of the scale values, $s' = as + b$, would reproduce the rank order of the matrix entries (Krantz et al., 1971).

In summary, the subtractive model predicts no interaction between row and column stimuli, assuming the judgment function is linear. The data should plot as parallel lines. The marginal means can be used to estimate scale values, which are unique to a linear transformation. However, if the judgment function is nonlinear, the model only predicts that it should be possible to rescale the data to parallelism.

Ordinal Indeterminacy

When the overt responses are only considered an ordinal scale (i.e., if J in Fig. 3 is only assumed to be strictly monotonic), it becomes more difficult to test models (Birnbaum, 1974a, 1974b; Birnbaum & Veit, 1974a, 1974b). If the data show ordinal violations of the theory, it is agreed that the model should be rejected. The difficulty occurs when the data are ordinally consistent with the model but metrically (numerically) inconsistent. When is it appropriate to transform data to fit the model and then conclude that the model fits? If the data can be transformed to fit the model, the scale values can be derived from the transformed data, and the inverse transformation could be interpreted as the judgment function (Birnbaum, 1974a, 1974b). However, such transformation may be theoretically inappropriate; the numerical deviations may represent "true" violations of the theory that should not be scaled away. For more extensive discussion of this problem including methods for dealing with it, see Birnbaum (1974a).

The case of ratio and subtractive models is an example of this problem. The ordinal requirements of the subtractive model and the ratio model are equivalent. Hence it is not possible to discriminate these models on the basis of ordinal information in a single experiment without some extra constraint. Data that are numerically consistent with the ratio model can be transformed to fit the subtractive model because $\log(R_{ij}) = \log(s_j/s_i) = \log s_j - \log s_i$. The center panel of Fig. 5 shows the results of a logarithmic transformation of the ratios in the left-hand panel of Fig. 5. Data that are consistent with the subtractive model can be exponentially transformed to fit the ratio model. For a single set of data, $s = \exp(s^*)$, where s is the scale value based on the ratio representation, and s^* is the scale for the subtractive representation.

If we assume that J is linear, then ratio and subtractive models can be distinguished on the basis of the metric properties of the raw (untransformed) data. However, if we do not assume that the judgment function is more than monotonic, we are forced to select a representation on the basis of some arbitrary criterion such as the task given the subjects or convenience. However, additional criteria can be specified to help resolve some of the indeterminancy.

SCALE CONVERGENCE CRITERION

By postulating that scales should be independent of the task, additional constraints are provided that limit the number of permissible transformations of the data. According to the stimulus scale convergence criterion, transformations of the data are deemed *appropriate* if they simultaneously fit models to data and lead to scales that agree. In this framework the scale attains greater status in that it becomes an intervening construct that can be used to account parsimoniously for an otherwise complicated set of relationships (Garner, Hake, & Eriksen, 1956; Krantz, 1972; Krantz et al., 1971; Anderson, 1972; Cliff, 1973; Birnbaum, 1974a; Birnbaum & Veit, 1974a; Shepard, 1976).

Although a single matrix of ordinally consistent data could be rescaled to fit either a ratio model or a subtractive model, Fig. 5 shows that two matrices with the assumption of scale invariance provide greater constraint: Ratios and differences of the *same* scale values are *not* monotonically related. For example, $7 - 5 > 2 - 1$ but $7/5 < 2/1$. The dashed lines in Fig. 5 connect pairs with equal differences of 1. Note that the ratios 2/1, 3/2, 4/3 are not equal but approach 1 as the constant interval (2-1, 3-2, 4-3) is moved up the scale (dashed curve in middle of Fig. 5). If there are both ratio and subtractive operations, then there will be two *different* rank orderings in the matrices. Since the scale that reproduces the order of the ratios is unique to a power function, and the subtractive scale is unique to a linear function, it follows that the common scale that reproduces the two different orders is unique to a similarity transformation. Thus, if the two orders are consistent with the models and interlocked by a common scale, the scale values constitute a ratio scale (Krantz et al., 1971).

On the other hand, if there is only one operation for both "ratios" and "differences," then both instructions will generate the same ordering of the pairs (Birnbaum & Veit, 1974a). If $R_{ij} = J_{\mathrm{R}}\,(s_j \odot s_i)$ and $D_{ij} = J_{\mathrm{D}}\,(s_j \odot s_i)$, where $\odot$ represents the comparison operation, then $s_j \odot s_i = J_{\mathrm{D}}^{\;-1}\,(D_{ij})$; hence, $R_{ij} = J_{\mathrm{R}}\,[J_{\mathrm{D}}^{\;-1}\,(D_{ij})]$. Since J_{R} and J_{D} are strictly monotonic, it follows that R_{ij} will be monotonically related to D_{ij} if there is only one operation.

In summary, there are two simple possibilities: (1) the two rank orders will be distinct, consistent with the respective models and appropriately interlocked by a common scale; (2) the rank order of the data in both matrices will be the same, consistent with the hypothesis that subjects perceive only a single comparison between a pair of stimuli. It is also possible that the data would be inconsistent with both of these alternatives, calling the models and/or the scale convergence criterion itself into question.

Empirical Evidence: One Operation

Figure 6 plots mean estimations of "ratios" of darkness, mean ratings of "differences," and rescaled values. The stimuli were those of Fig. 4, which were administered to 44 undergraduates at the University of Illinois. Half the subjects performed either task first, with no evidence of task order effects.

The left-hand panel of Fig. 6 shows mean "ratio" estimations plotted against marginal means for the column stimulus. The mean estimations (open circles) come very close to the bilinear pattern (lines) predicted by the ratio model.

The right-hand panel plots the mean "difference" estimations against the column marginal means. The data appear nearly parallel, as predicted by the subtractive model. Considering the fit of the raw numerical data to the models implied by the tasks, it would be tempting to conclude that subjects are actually

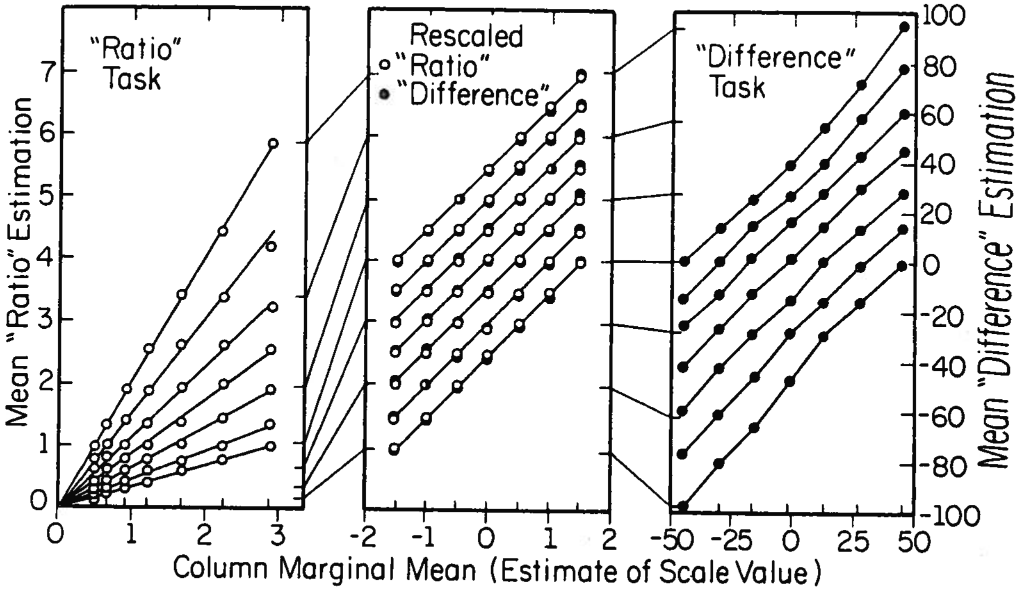

FIG. 6 Mean "ratios" and "differences" of darkness of stimuli in Fig. 4, plotted as in Fig. 5. Although "ratio" task data fit ratio model, and "difference" task data fit subtractive model, both sets of data have equivalent rank orders. Center panel shows data rescaled to parallelism. Transformations are represented by lines connecting panels. Note that the data are *not* like the predictions shown in Fig. 5, since both sets of rescaled values coincide. (From Birnbaum and Stegner, 1976.)

computing ratios for "ratios" and differences for "differences." It is shown later, however, that this interpretation leads to contradiction of the scale convergence criterion.

The center panel shows the results of separate rescalings of both sets of data to parallelism via MONANOVA (Kruskal & Carmone, 1969), a computer program that seeks a monotonic transformation to reduce interactions in analysis of variance. Both the "ratio" task data (open circles) and the "difference" task data (solid points) appear to coincide after transformation. The two tasks seem to generate a single order that can be represented by either a ratio model, or a subtractive model, but not both.

It may seem surprising that subjects given two different tasks provide numbers for the respective matrices that do not obey the ordinal requirements of two operations on one scale. If the subjects had covertly assigned numbers to the stimuli and then calculated ratios and differences on these covert numbers, the results would have been quite different, because the two matrices would have obeyed the predictions of two models on a common scale. Instead, if we accept the premise that the scale values of darkness are independent of the task, it appears that the comparison process is also independent of the task to judge "ratios" or "differences."

More Evidence

The results of this demonstration experiment are typical of results obtained with judgments of the heaviness of lifted weights (Birnbaum & Veit, 1974a), numerical magnitude (Rose & Birnbaum, 1975; Birnbaum, 1974b), shades of gray (Veit, 1974), likeableness of adjectives (Hagerty & Birnbaum, 1976), and loudness (Birnbaum & Elmasian, 1977).

Birnbaum and Elmasian (1977) presented pairs of 1000-Hz tones varying in sound pressure level and asked subjects to compare the loudness of the two tones. The pairs were constructed from a 5 X 9, B X A, factorial design in which the first tone (B) varied from 42- to 90-dB SPL in 12-dB steps; the second tone (A) covered the same range in 6-dB steps. Each subject served in four daily sessions, two for "ratios" and two for "differences," completing 10 replications of the design per session. Separate analyses, performed on the data for each subject-day, led to the conclusion that estimates of "ratios" and ratings of "differences" are each roughly numerically consistent with their respective models. The mean "ratio" estimations, shown in Fig. 7, plotted as in Fig. 6, are nearly bilinear. The mean "difference" ratings (9-point scale), shown in Fig. 8, are nearly parallel. However, the two orders for each subject are approximately the same for both tasks and can therefore be represented by a single comparison operation.

The data for both tasks were transformed to parallelism. Figure 9(A) shows the predicted results for the transformed scores, based on the theory that the subjects can compute both ratios and differences of loudness. In both panels, solid points connected by straight lines represent rescaled "differences," open

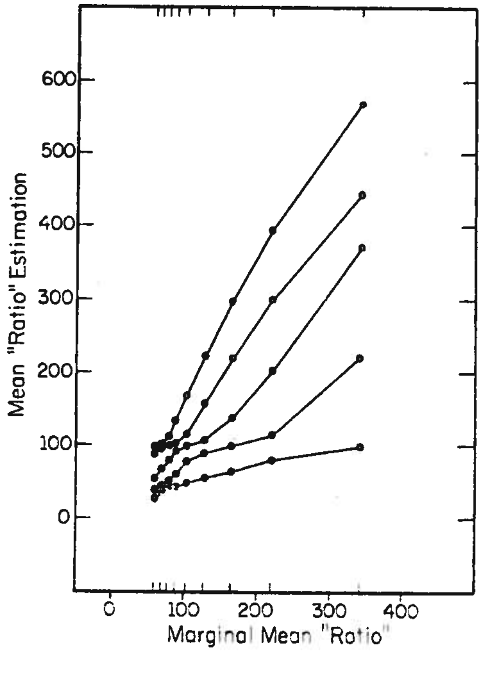

FIG. 7 Mean estimation of "ratio" of loudness, plotted as in left panel of Fig. 6. Modulus was 100. (From Birnbaum and Elmasian, 1977.)

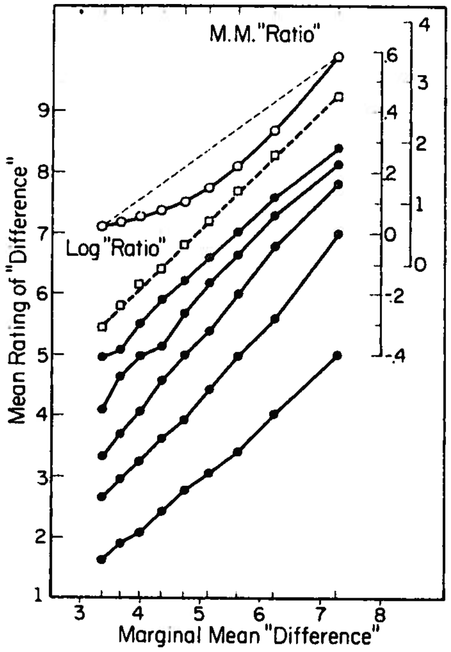

FIG. 8 Mean rating of "difference," as in the right panel of Fig. 6 (solid points). Open circles plot marginal mean "ratios" as a function of marginal mean "differences." Open squares plot marginal mean log ("ratio") against marginal mean "difference." Mean "ratios" and log "ratios" are to be read against far-right and right ordinates, respectively. Linearity of open squares agrees with the theory that one operation underlies both tasks and that magnitude estimations are an approximately exponential function of ratings. (From Birnbaum and Elmasian, 1977.)

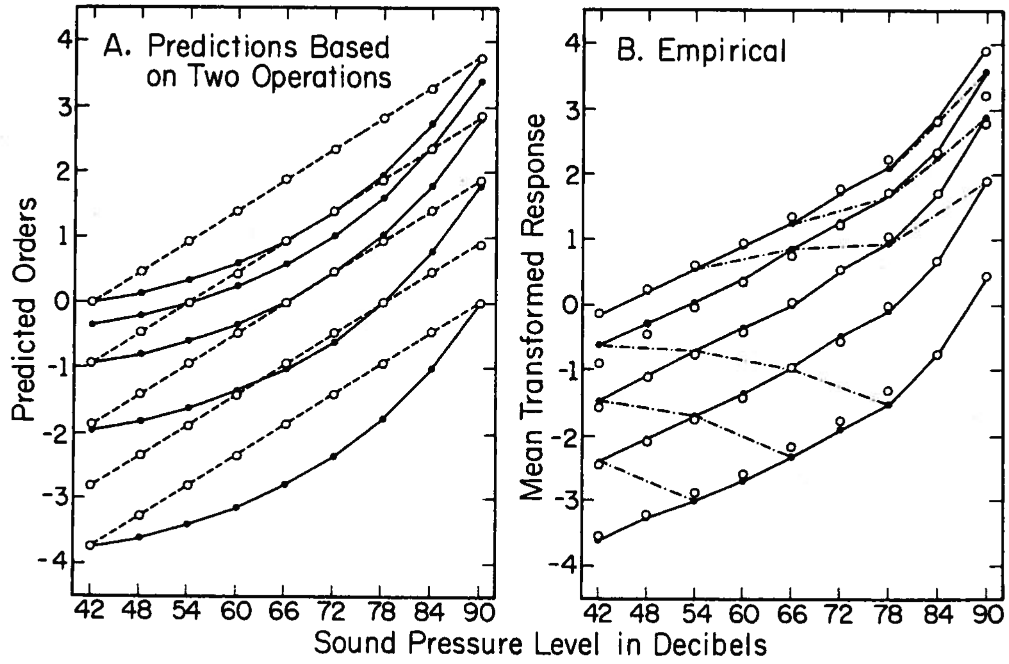

FIG. 9 Transformed "ratios" and "differences" of loudness. (A) Left-hand panel shows predicted transformed ratios (open circles) and differences (solid points) assuming power functions for loudness. Hypothetical values were computed by taking ratios and differences of power functions of physical sound pressure and then transforming to parallelism on a common scale. (B) Right-hand panel plots actual mean transformed values as a function of the sound pressure level of tone *A* with a separate curve for each level of tone *B*, which varied from 42 dB to 90 dB in 12-dB steps. Solid points connected by solid lines represent transformed "differences"; open circles are transformed "ratios." Broken lines connect pairs of equal physical ratios; lowest broken line is for –36 dB, next is for –24, etc. (From Birnbaum and Elmasian, 1977.)

circles represent rescaled "ratios." Although the particular hypothetical predictions shown in Fig. 9(A) were computed using scale values for both tasks that were power functions of physical intensity, the general relationship between the solid and dashed curves would remain the same for other scale values (transformation of the abscissa). Figure 9(A) shows the relationship between transformed ratios and differences, showing again how different the results should be if subjects actually used two operations.

Figure 9(B) shows the actual mean transformed scores, plotted as a function of sound pressure level for the second stimulus (A) with a separate curve for each level of the first stimulus (B). The mean transformed scores are representative of the single subject data presented by Birnbaum and Elmasian (1977). Each set of rescaled data is nearly parallel, and the two sets are nearly identical. The similarity of the orders can be seen by the coincidence of circles (representing rescaled "ratios") and lines (connecting rescaled "differences"). These results are consistent with the hypothesis that there is but one loudness comparison for "ratios" and "differences."

The broken curves in Fig. 9(B) connect pairs with equal physical ratios. Assuming a ratio model, the power function ($s = a\Phi^b$) implies that equal physical ratios should receive equal "ratio" judgments, since $R_{ij} = J_{\mathrm{R}}\ (s_j/s_i) = J_{\mathrm{R}}\ (\Phi_j{}^b/\Phi_i{}^b) = J_{\mathrm{R}}\ [(\Phi_j/\Phi_i)^b]$. Instead, the broken curves of Fig. 9(B) show that equal physical ratios receive more extreme judgments at the upper end of the scale. Scale values for the ratio model are inconsistent with a power function for loudness as a function of physical intensity.

Theories We Can Reject

The finding that instructions to judge "differences" and "ratios" lead to the same ordering of stimulus pairs allows us to reject the theory that subjects use two operations on a common scale. That is, for the continua and conditions studied in our experiments, we reject the theory that can be explicated in the following four premises:

P_1 (independence): The scale value of a stimulus is independent of the stimulus with which it is compared.
P_2 (scale convergence): $s^* = s$; the scale value of a stimulus is independent of instructions to judge "ratios" or "differences."
P_3 (ratio model): $R_{ij} = J_{\mathrm{R}}\ (s_j/s_i)$
P_4 (subtractive model): $D_{ij} = j_{\mathrm{D}}\ (s_j{}^* - s_i{}^*)$

The first premise explicitly assumes that the scale value of row or column stimulus is independent of the stimulus with which it is paired. The scale values are assumed to be independent of the task in premise P_2. In premises P_3 and P_4, which state the models, J_R and J_D are any monotonic functions.

The finding that both models fit the data seems inconsistent with previous hypotheses that the reason that "ratio" and "interval" techniques yield different scales is due to a computation error. For example, Stevens (1971) contended that "the human being, despite his great versatility, has a limited ability to effect linear partitions on prothetic continua." The contention that subjects make a miscalculation in computing differences leads to violations of the subtractive model. The present data show that whatever subjects are doing for "intervals," they order the pairs in the same way for "ratios." There seems no evidence to support the contention of two operations with a distortion or bias in one computation, since both sets of data satisfy the ordinal requirements of the subtractive (or ratio) model of comparison.

The data do not support the theory that there are two operations with two scales of sensation, one for "ratios" and another for "intervals," both of which are power functions of physical intensity (Marks, 1974; Stevens, 1971). If $s = \Phi^b$ and $s^* = \Phi^c$, then $s^* = s^{c/b}$. Therefore, the two scales would be related by a power function. This theory turns out to be equivalent in its ordinal predictions to the theory that there are two operations on one scale. In fact, premise P_2 could be replaced by $s^* = as^{\beta} + \gamma$, and the theory would still predict different orders for the two tasks.

The data do not require more complicated theories, for example, that there is one operation with two scales of sensation or that neither model could provide a representation of either set of data.

Theories We Can Entertain

It would be possible to retain the previous premises P_1, P_3, and P_4 on the basis of the present data if the scale convergence premise (P_2) were replaced with the premise, $s = \exp(s^*)$. However, to retain the scale convergence criterion requires that the other premises be modified, because the entire theory ($P_1 - P_4$) cannot account for the data. There are two simple theories that retain scale convergence and "save the phenomena" (reproduce the data of these experiments).

1. *Ratio theory:*
   - P_1: (independence)
   - P_2: (scale convergence): $s^* = s$
   - P_3: (ratio model): $R_{ij} = J_R\ (s_j/s_i)$
   - P_5: (ratio model): $D_{ij} = J_D\ (s_j^*/s_i^*)$

This theory asserts that subjects compare stimuli by computing ratios, irrespective of the instructed task. The data for the darkness of the dot patterns (Fig. 6) imply that J_R is at least of the form $R_{ij} = a\,(s_j/s_i)^b + c$, since the data are very nearly bilinear. The judgment function for ratings of "differences," J_D, would have to be approximately logarithmic to account for the near-parallelism of the "difference" ratings.

According to the ratio theory, scale values estimated from the column marginal means, $\bar{R}_{.j}$, are 50, 67, 92, 122, 167, 223, and 287 for the seven levels of darkness. These scale values can be well approximated as a power function of the number of dots, $s = \Phi^{.72}$. Since scale values defined by the ratio model are unique only to a power scale, the exponent would have little meaning unless it were assumed that J_R were linear.

2. *Subtractive theory:* This theory asserts that the basic comparison is a difference, irrespective of task.

P_1: (independence)
P_2: (scale convergence): $s^* = s$
P_6: (subtractive): $R_{ij} = J_R\,(s_j - s_i)$
P_4: (subtractive): $D_{ij} = J_D\,(s_j^* - s_i^*)$

According to subtraction theory, the data for the darkness example imply that J_R must be nearly exponential, since the "ratio" estimations (left panel of Fig. 6) are nearly bilinear. If $R_{ij} = \exp(s_j - s_i)$, it follows that a subtractive operation would lead to data satisfying the predictions of the ratio model, because $\exp(s_j - s_i) = \exp(s_j)/\exp(s_i)$.

The scale values for the subtractive model, estimated after transforming the data for both tasks to parallelism, are: -1.48, -1.00, $-.50$, $-.02$, $.50$, 1.02, and 1.50. These scale values are very nearly equally spaced, as are the logarithms of the number of dots, indicating that the scale values could be well approximated by a logarithmic psychophysical function of the number of dots, $s = \log \Phi$.

Summary and conclusions. The data are compatible with the theory that the same comparison operation applies for both tasks, since the two orderings are equivalent. If the operation is represented by a ratio model, then the scale values for darkness can be fit as a power function of number of dots. Furthermore, the ratio interpretation implies that the judgmental transformation for magnitude estimation must be at least a power function and perhaps even linear. To explain the near-parallelism for the difference task would require that the J_D function be approximately logarithmic. If the operation is represented by a subtractive model, however, then the judgmental transformation for "ratios," J_R, would be exponential, and J_D would be approximately linear. The psychophysical function would be well approximated by a logarithmic function. Hence the conclusions for the stimulus and response scales derived from the data depend on the model or theory that is assumed.

The simplest interpretation appears to be that there is only one comparison operation, only one psychophysical scale, and two different judgmental transformations depending on whether the response is a category rating or magnitude estimation. This interpretation is consistent with Torgerson's (1961) theory and is consistent with the data of numerous experiments using factorial designs in which subtractive and ratio models can be evaluated in tandem. If the operation is a ratio, the J_R transformation for magnitude estimation would receive support from the fact that the data are numerically consistent with the model, but the J_D transformation for ratings would be near-logarithmic. On the other hand, if the operation is subtraction, J_D would be nearly linear and J_R exponential.

The inferred psychophysical scales, s, psychophysical function, H, and judgmental transformations, J, all depend on the assumed representation. Does it make sense to ask which operation is "really" correct? Torgerson (1961) noted that this question can not be resolved in the two-stimulus case. The next section discusses a nonmetric four-stimulus approach in which this question becomes meaningful in the sense that experiments could refute one theory or the other.

Scale and Theory Dependence

A simplistic view of functional measurement maintains that the metric fit of a model simultaneously "validates" the model and the response scale. Had the data for only one task (*either* the "difference" or "ratio") been obtained, it might have been tempting to conclude that the fit of the model "validates" both model and scale. However, the scale convergence criterion combined with the data for the two tasks implies, in spite of the metric fit of both models to the raw data, that at least one of the models and one of the scales must be rejected. The present findings show that extreme caution must be used in interpreting the metric fit of a model as evidence for "validity."

Birnbaum and Veit (1974b) introduced the term *scale-dependent* to refer to research in which the determination of the "appropriate" model depends on the arbitrary choice of the "valid" dependent variable, and the "validity" of the response procedure circularly depends on the arbitrary choice of model. For example, if the validity of ratings were assumed for "difference" judgments, the subtractive model would be chosen; the ratio model and magnitude estimation would be rejected. On the other hand, if the validity of magnitude estimations were assumed, then the ratio model would be preferred, and it would be concluded that ratings and the subtractive model are not valid. Thus the choice of model depends on the choice of dependent variable and vice versa.

In scale-dependent tests of the size-weight illusion, Anderson (1972) fit an additive model to rating data; however, J. C. Stevens and Rubin (1970) and Sjoberg (1969) fit ratio models using magnitude estimations. Sarris and Heineken (1976) replicated these results for the size-weight illusion in a single experiment in which only the change of dependent variable sufficed to change the data from parallel to bilinear. Weiss (1972) found that ratings of average darkness of a pair of gray chips were almost consistent with a constant-weight averaging model (additive), whereas magnitude estimations were nearly consistent with a geometric averaging model (multiplicative). Birnbaum and Veit (1974b) have noted that if the response procedure only affects the judgment function, J, and if magnitude estimations are exponentially related to ratings, it follows that if ratings fit an additive (or subtractive) model, magnitude estimations would be expected to fit a multiplicative (or ratio) model.

These scale-dependent experiments, when analyzed in conjunction with the scale-convergence criterion, appear consistent with the proposition that the operations are unaffected by the response procedure, but the judgmental transformation depends on whether category ratings or magnitude estimations are used as the dependent variable. These experiments also illustrate that tests of internal consistency and certain types of "cross-task validation" (e.g., Anderson, 1972), in which the *same* model is applied for both tasks, are not diagnostic tests of the "validity" of the models, scales, or response procedures, since choice of a different dependent variable can alter the apparent form of both models while still retaining cross-task scale convergence.

SCALE-FREE TESTS

The scale-free approach requires only the ordinal information in the data, plus some theoretical assumptions, to test models with far greater constraint than has been achieved in the past. The scale-free approach was developed by Birnbaum (1974a, experiment IV) to test the additive and constant-weight averaging models of impression formation. Birnbaum and Veit (1974b) have applied it to the size-weight illusion, and Veit (1974) has employed a novel application of the technique to the ratio-difference problem. The following subsections expand on the work of Veit (1974, in press) and describe a recent experiment by Hagerty and Birnbaum (1976) that illustrates the scale-free approach for the ratio-difference problem.

Quantitative Relations for Pairs

Suppose for the moment that the "true" operation is subtraction. This could be so for two distinct reasons: (1) it may be that for some reason subjects have only one operation for comparing two stimuli—perhaps, metaphorically, they do not have the neural circuitry to do anything else; and (2) on the other hand, it may be that the operation employed depends on the internal stimulus representation. Perhaps the sensation values should be represented by points on a line with arbitrary origin (i.e., places, not lengths). In such a representation, distances or differences are sensible, but ratios are not. Thus, when asked to judge "ratios," the subject may actually judge differences (Birnbaum & Veit, 1974a).

If the second interpretation were correct, then intervals (differences) would have a well-defined zero point even if the stimulus values did not. Hence subjects might be able to judge both ratios and differences of *pair intervals*, even if they could not judge both ratios and differences of stimulus magnitudes (Veit, 1974). On the other hand, if the interpretation that there is only one operation for such judgments were the case, then only one operation would be observed for comparisons of pairs.

In order to achieve tests of these possibilities, it is necessary to employ tasks in which the subject receives four stimuli on each trial and is asked to compare two pair relations. The four tasks discussed below are "ratio of ratios," "ratio of differences," "difference of ratios," and "difference of differences." The next subsection outlines four models corresponding to these four tasks.

Four Models of Comparison of Pairs

Ratio of ratios model. This model can be written:

$$RR_{ijkl} = J_{\mathrm{RR}}\ [(s_j/s_i)\ /\ (s_l/s_k)]\,, \tag{8}$$

where RR_{ijkl} is the "ratio of ratios" estimation of the ratio of stimulus levels j to i, relative to the ratio of stimulus levels l to k; J_{RR} is the monotonic judgmental transformation from impressions to overt responses; s_j, s_i, s_l, and s_k are the scale values of the four stimuli, factors A, B, C, and D, respectively, in a four-way factorial design.

Predictions for the ratio of ratios model are shown in Fig. 10, where the calculations are based on scale values of successive integers from 1 to 7. The experimental design portrayed in the upper left of Fig. 10 is a 7 X 7 X 3 design in which the numerator pair is composed of a 7 X 7, A X B, factorial design, and there are three levels of the divisor ratio, *C/D*, as shown in the figure.

The model predicts a trilinear interaction, in which the bilinear $A \times B$ interaction is multiplied by the effect of the divisor ratio. Hence, each $A \times B$ interaction should be similar, with greater divergence for smaller divisor ratios. It should be noted that since $J_{\mathrm{RR}}^{-1}\ (RR_{ijkl}) = (s_j/s_i)\ /\ (s_l/s_k)$, $\log\ [J_{\mathrm{RR}}^{-1}\ (RR_{ijkl})] = \log s_j - \log s_i - \log s_l + \log s_k$; hence this model is ordinally additive in form.

Ratio of differences model. This model can be written:

$$RD_{ijkl} = J_{\mathrm{RD}}\ [(s_j - s_i)\ /\ (s_l - s_k)]\,, \tag{9}$$

where RD_{ijkl} is the judged "ratio of differences" of the difference between stimuli s_j and s_i relative to the difference between s_l and s_k; J_{RD} is the judgmental transtormation.

Metric predictions for this model are shown in the lower left panel of Fig. 10. Since the numerator contains a subtractive model, the curves for the levels of

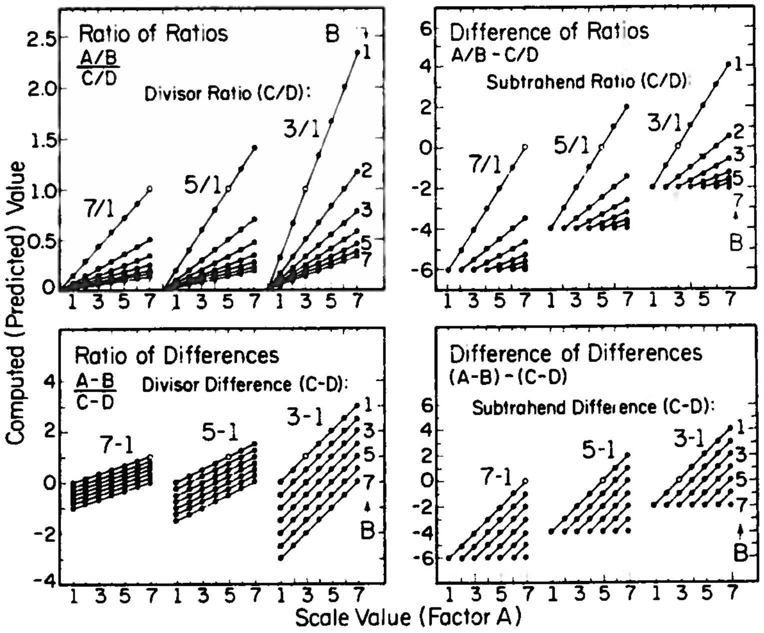

FIG. 10 Predictions for four polynomial theories of comparison of pairs. Successive integers from 1 to 7 are used as scale values for factors A and B. Separate fans of curves are shown for the three divisor or subtrahend pairs (open circles). The two "difference" tasks show predictions only for upper triangular design, in which A-B differences are positive.

$A \times B$ factors $(s_j - s_i)$ are parallel, indicating no interaction between A and B. However, the reciprocal of the divisor difference multiplies this numerator difference. Hence the smaller the divisor difference, the larger the vertical spreads and slopes of the curves. Therefore, the model predicts a bilinear interaction between the numerator and divisor difference. This model is a type of distributive model (Krantz & Tversky, 1971), in which the two dividend factors ($A \times B$) are jointly independent of the divisor (i.e., the ordering in the $A \times B$ matrix for each level of CD is the same), but the other pairs of factors are not (i.e., the ordering of the $A \times$ Divisor matrix depends on the level of B). Hence this model can be distinguished from the others on the basis of ordinal properties of the data. For more detailed discussion of diagnostic analyses for the ratio of differences model, see Veit (1974). Conjoint measurement analyses of general classes of polynomials are discussed by Krantz et al. (1971) and Krantz and Tversky (1971).

Difference of ratios model. This model can be written:

$$DR_{ijkl} = J_{\mathrm{DR}}\ [(s_j/s_i) - (s_l/s_k)], \tag{10}$$

where DR_{ijkl} is the rating of the "difference between two ratios," and J_{DR} is the judgmental transformation for this task.

Predictions for this model are shown in the upper right panel of Fig. 10 for the positive triangle of the 7×7 design (where $s_j - s_i \geqslant 0$), with a separate plot for each subtrahend ratio (s_l/s_k). The predicted pattern is a bilinear $A \times B$ fan for each subtrahend. The bilinear fans for different subtrahends should be congruent, differing only by an additive constant. Thus there is a bilinear interaction between factors A and B, but factors A and B should not interact with the subtrahend pair (CD). This model is a dual-distributive polynomial that can be distinguished from the others on the basis of ordinal information in the data (Krantz & Tversky, 1971).

Difference of differences. This model can be written:

$$DD_{ijkl} = J_{\mathrm{DD}}\,[(s_j - s_i) - (s_l - s_k)]\,, \qquad (11)$$

where DD_{ijkl} is the rating of the "difference between two differences," and J_{DD} is the judgmental transformation.

The lower right panel of Fig. 10 shows predictions for the positive upper triangle of the 7×7 design (where $s_j - s_i \geqslant 0$). The predicted pattern is one of three-way additivity. Each set of curves is parallel, differing only in an additive constant for each subtrahend pair. Note that although this model is also formally additive and therefore equivalent in an ordinal sense to the ratio of ratios model, scale convergence provides an additional constraint that implies different orderings when the data for the two tasks are compared.

Possible Outcomes

The constraints of the four-stimulus models are such that data can be diagnosed as additive, distributive, or dual-distributive [Eqs. (8) and (11), (9), or (10), respectively] on the basis of the ordinal properties of the data. Scale convergence among the four sets of data (and with the scales from the two-stimulus experiments) provides additional constraints on possible solutions so that different potential outcomes can be distinguished on the basis of the data. Five of these possible outcomes deserve closer attention.

One operation. One simple possibility is that there is but one operation by which either a pair of stimuli are compared or a pair of stimulus pairs are compared. If this were the case, then the data for all of the four-stimulus tasks could be rescaled to fit either the ratio of ratios model or the difference of differences model. If all of the four-stimulus tasks resulted in the same ordering, the single comparison operation would remain indeterminate.

Subtraction theory. A second possibility is that the basic operation by which two stimuli are compared is subtraction (Birnbaum & Veit, 1974a; Veit, 1974). This could occur if the subjective stimulus representation is an interval scale, like points along a line with undefined origin. The apparent fit of the ratio model for the "ratio" task would be accounted for by postulating that the judgmental transformation for magnitude estimation of "ratios" is exponential. This theory suggests that subjects could judge ratios of intervals, since differences have a well-defined zero point even when the stimuli do not. Data for the other four-stimulus tasks might be expected to be consistent with the difference of differences model. The "ratio of ratios" tasks would be expected to require a logarithmic transformation to be fit to the difference of differences model, since the dependent variable, magnitude estimation, is presumed to be an exponential function of subjective value. The scales derived from these models would then agree with the subtractive theory of the pair tasks.

Ratio theory. The basic operation could be a ratio, upon which either differences or ratios could be judged. If this theory were the case, than all of the data could be fit to the ratio and ratio of ratios models, except for the "difference of ratios" task; this task should fit the difference of ratios model, defining scales that agree with the ratio interpretation of the other tasks.

Comparison of pairs validity. It may be that the data for all of the four-stimulus tasks fit their respective models with a single underlying scale as in Fig. 10. If this outcome were obtained, the common scale could be used to decide between the ratio and subtractive theories for the two-stimulus tasks.

Two worlds. Perhaps there are two subjective "worlds," one for each operation, with an exponential relationship between the scales. Perhaps the "ratio," "ratio of ratios," and "difference of ratios" data would fit their respective models with one scale. But suppose the "difference," "difference of differences," and "ratio of differences" tasks could be fit to their respective models with another scale that differed from the first. This "impossible figure" outcome, which would look consistent within each realm but inconsistent between realms, could come about if the subject assigned numbers to each *pair* and computed on the numbers to compare pairs.

Evidence for Subtraction Theory

Shades of gray. Veit (1974, Experiment I) found that ratings of "differences" and estimations of "ratios" of darkness of gray papers were monotonically related, consistent with the theory that one operation applies to both tasks. In a second experiment, she found that magnitude estimations of "differences" generate the same ordering of pairs as magnitude estimations of "ratios" and category ratings of "intervals." However, magnitude estimations of "differences" showed a divergent interaction that required rescaling (interpreted as J^{-1}) to render the curves parallel. This finding is consistent with the interpretation that the comparison operation and scales are independent of the response procedure but that the choice of response procedure affects the judgmental transformation. Beck and Shaw (1967) reached similar conclusions for magnitude estimations of loudness intervals.

In her third experiment, Veit (1974) introduced the ratio of differences model as a test of the subtractive representation. Figure 11 plots the magnitude estimations of "ratios of differences" as in the lower left of Fig. 10, except that larger divisor differences are on the right. The data clearly show the pattern predicted by the ratio of differences model. Consistent with the model, Fig. 12 shows that the small $A \times B$ interactions seen in Fig. 11 can be removed by separate transformation of the data for each divisor difference. It was also possible to fit the ratio model to the numerator/demoninator $[(A - B) / (C - D)]$ comparisons. However, as predicted by the ratio of differences model, it was not possible to eliminate interactions between A or B and the divisor difference. Tests of joint independence (see Krantz et al., 1971) for individual subjects were consistent with the interpretation of the ratio of differences model and inconsistent with the other simple models.

The scale values for the seven levels of reflectance derived from the ratio of differences model were used to evaluate alternative theories for the two-stimulus judgments. These scale values were consistent with the subtractive representation of the simple "difference" and "ratio" tasks. The ratio theory implies scale values that contradict the ratio of differences model. Veit (1974, in press) noted that this finding makes the ratio interpretation implausible. In order to represent

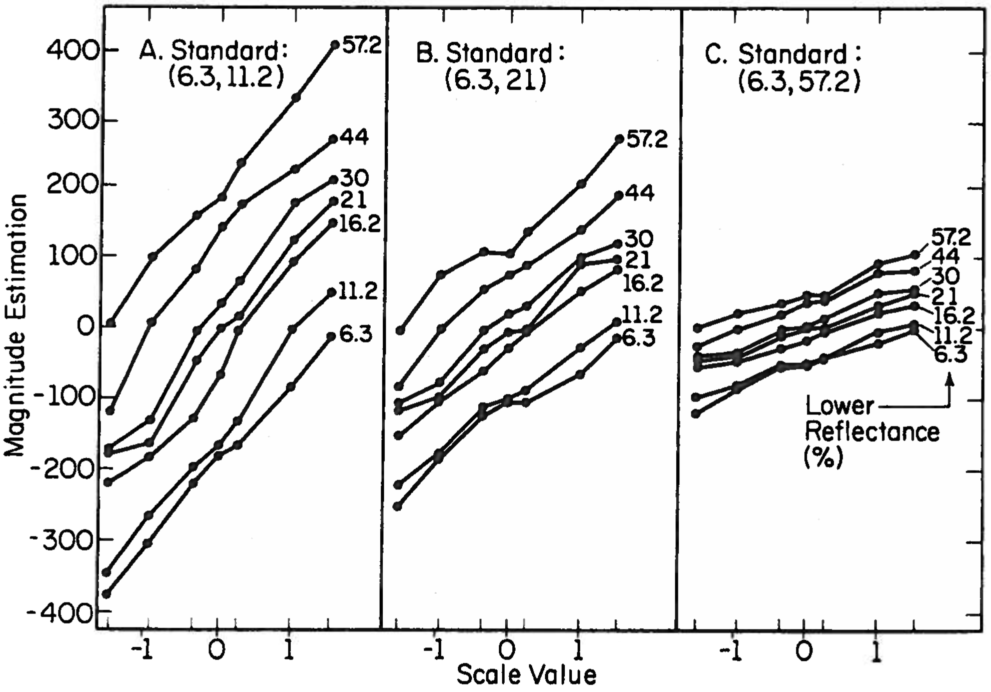

FIG. 11 Mean estimations of "ratios of differences." Each panel shows results for a different divisor difference (C-D), labeled "standard" in the figure. Curve parameters are reflectance values for subtrahend (factor B); abscissa spacing represents estimated scale values for minuend stimulus (factor A). (From Veit, 1974.)

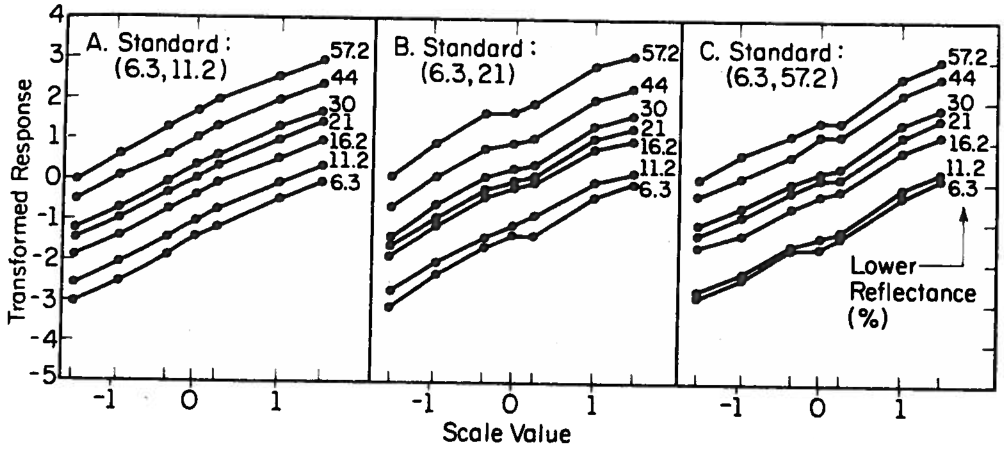

FIG. 12 Transformed mean response as a function of Factor *A* scale value (minuend) with a separate curve for each level of factor *B* (subtrahend). A separate rescaling was performed for each divisor (standard) difference. (From Veit, 1974.)

the numerator "difference" with a ratio model and preserve scale convergence, exponential transformation of Eq. (9) represents "ratios of differences" by the model,

$$\begin{aligned} RD_{ijkl} &= \delta_{ij} J_{\mathrm{RD}} \{\exp\,[(s_j - s_i)/(s_l - s_k)]\} \qquad (12) \\ &= \delta_{ij} J_{\mathrm{RD}} \{[\exp(s_j - s_i)]^{1/(s_l - s_k)}\} \end{aligned}$$

hence

$$RD_{ijkl} = \delta_{ij} J_{\mathrm{RD}} \{(s_j^*/s_i^*)^{1/(s_l - s_k)}\} \qquad (13)$$

where $s^* = \exp(s)$, and δ_{ij} is the sign of $s_j - s_i$. The ratio interpretation requires two different models for "differences," s_j^*/s_i^* and $s_l - s_k$ (for numerator and denominator "differences," respectively), with two different scales, s^* and s. Furthermore, an instructed "ratio" would be represented by two different models: a ratio model for "ratios" of two stimuli and an exponential-power function for "ratios of differences."

Because of these complexities, Veit (1974) rejected the ratio theory as a viable alternative. Scale values derived from the subtractive representation of magnitude estimations of "ratios," ratings of "differences," magnitude estimations of "differences" and "ratios of differences" were all in close agreement, consistent with the simpler interpretation of the subtractive theory (Veit, 1974; in press).

The possibility remains, however, that the difference of ratios model would fit data obtained in a "difference of ratios" task, yielding scales that would agree with the ratio interpretation of the "ratio" and difference," two-stimulus tasks. This "two worlds" potentiality and others were checked by employing all of the four-stimulus tasks in an experiment by Hagerty and Birnbaum (1976.)

Likeableness of adjectives. Hagerty and Birnbaum (1976) studied judgments of the likeableness of hypothetical persons described by adjectives. For example, how much more would you like someone who is *sincere* than someone who is *mean?* Six tasks were employed, including both of the two-stimulus tasks and all of the four-stimulus tasks. The same subjects performed several tasks the first day for practice, then returned for two more days to complete all of the tasks. Adjectives were chosen on the basis of normative ratings to represent seven levels of likeableness, for example: *cruel, irritating, clumsy, hesitant, thrifty, capable,* and *sincere.*

Data for the "ratio" and "difference" tasks are shown in Fig. 13. The factorial stimulus design was a 4 X 7, *B* X *A*, using different adjectives for the two factors. In the left panel of Fig. 13, the "ratio" estimations (with a modulus of 100) show the approximate bilinear form of diverging curves when plotted against the marginal means. The right panel of Fig. 13 shows that the "difference" ratings (on a 9-point scale) are approximately parallel. Both tasks yield data that can be rescaled to approximate parallelism as shown in the center panel of Fig. 13. The rescaled "ratios" (circles) and "differences" (points) are nearly identical, consistent with previous results and the interpretation that one comparison operation underlies both tasks.

The assumption of the ratio model and the linearity of magnitude estimations would imply that the marginal means (abscissa spacing of Fig. 13) represent a scale of likeableness of the adjectives. The interval between the lowest and middle adjectives *hesitant-cruel* is less than the interval between the two highest adjectives *sincere-capable.* The subtractive representation (abscissa spacing in center panel) leads to the opposite conclusion: The *hesistant-cruel* interval is the larger.

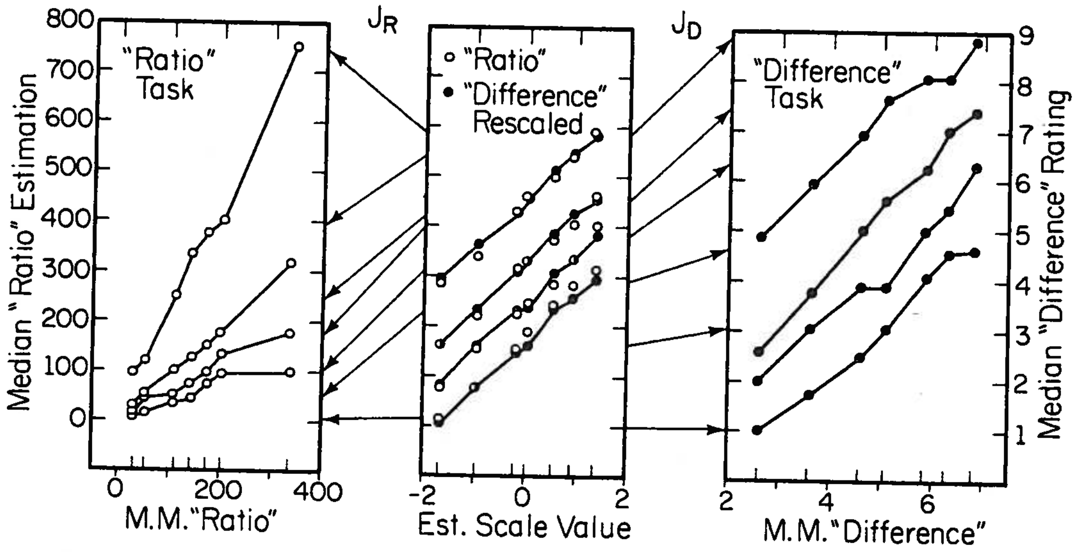

FIG. 13 "Ratios" and "differences" in likeableness of adjectives, plotted as in Fig. 6. *Left:* Shows median "ratios." *Center:* Shows that rank orders for both tasks are nearly identical and that rescaled data are roughly parallel. Assuming a subtractive model for both tasks, the transformations to overt responses (arrows) represent judgmental functions. *Right:* Shows median "differences." (From Hagerty and Birnbaum, 1976.)

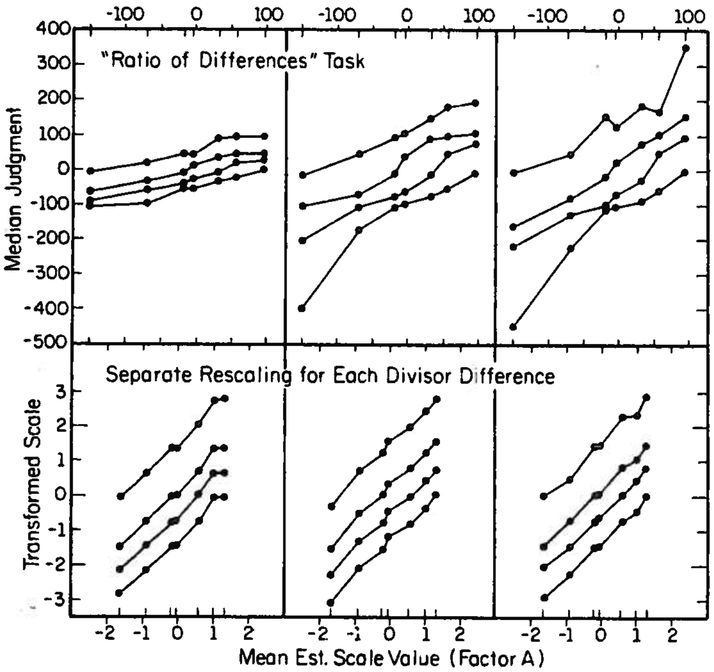

FIG. 14 "Ratios of differences" in likeableness. Upper panels plot median estimates as in Fig. 11. Lower panels plot rescaled values as in Fig. 12. Data are compatible with a ratio of differences model. (From Hagerty and Birnbaum, 1976.)

For the four-stimulus tasks, the 4 X 7 design was combined with three divisor or subtrahend pairs (*truthful-phony*, *truthful-listless*, and *practical-listless*). Figure 14 presents the results for the "ratio of differences" task plotted as in Figs. 10 and 11. Data for the largest divisor difference (truthful-phony) are on the left. The curves in Fig. 14 show the form predicted by the ratio of differences model (lower-left of Fig. 10): The smaller the divisor difference, the larger the slopes of the curves and vertical spreads between the curves. The lower panels plot transformed medians, showing that for each divisor, the data can be separately rescaled to parallelism. Other ordinal tests also indicated that the data could be represented by Eq. (9), and that monotonic transformation could not fit the data to any of the other models.

In sum, the "ratios of differences" in likeableness are consistent with the ratio of differences model, in agreement with the findings of Veit (1974).

Results for the "ratio of ratios" task are shown in Fig. 15, plotted as in Fig. 10. The median estimations, plotted in the upper panel as a function of marginal means, show the approximate trilinear divergent interactions anticipated by the ratio of ratios model. The lower panels show the rescaled medians (following monotonic transformation to fit the difference of differences model). Parallelism, linearity, and congruence of the three sets of curves would be evidence that a difference of differences (or ratio of ratios) model is ordinally compatible with the data. In spite of some deviations, the data appear in approximate agreement with the model.

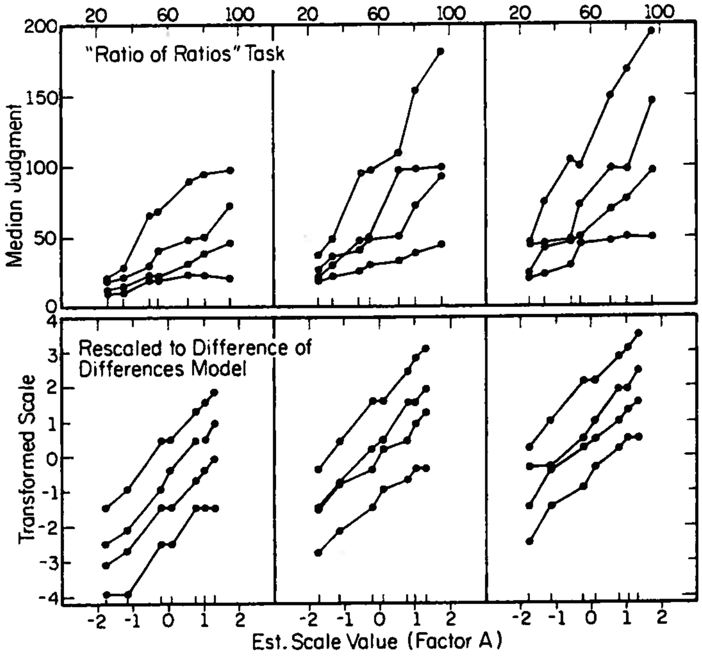

FIG. 15 Results of "ratios of ratios" task for likeableness judgments, plotted as in Fig. 10. Lower panels plot rescaled medians, fit to the difference of differences model. Scale values for difference of differences model agree with scale values for ratio of differences model fit in Fig. 14. (From Hagerty and Birnbaum, 1976.)

Figure 16 shows median ratings of "difference of differences" in likeableness for that portion of the 4 X 7 design in which the column adjective was rated on the average more likeable than the row adjective. Data have been rescaled to fit the difference of differences model; rescaled medians are plotted against estimated scale values in the lower panels. The near-linearity, -parallelism, and -congruence of the sets of curves is consistent with the predictions (Fig. 10) of the difference of differenres model.

Median ratings of "difference of ratios" are shown in Fig. 17, plotted as in Fig. 16. The data do not conform to the predictions of the difference of ratios model (see Fig. 10), which predicts diverging fans for each set of curves, nor could the data be transformed to fit the difference of ratios model. Instead, the data are very similar to the data for "difference of differences" (Fig. 16) and can be rescaled to fit the same model, yielding transformed values (lower panel of Fig. 17) that are nearly congruent with transformed values in Fig. 16. It thus appears that the complicated "two worlds" outcome did not materialize, since the "difference of ratios" task can be represented by a difference of differences model.

The scale convergence criterion can be used to select a set of representations for all of the tasks that give a unified picture of these data. Figure 18 provides a summary of tests of scale convergence for the simplest interpretation of the data.

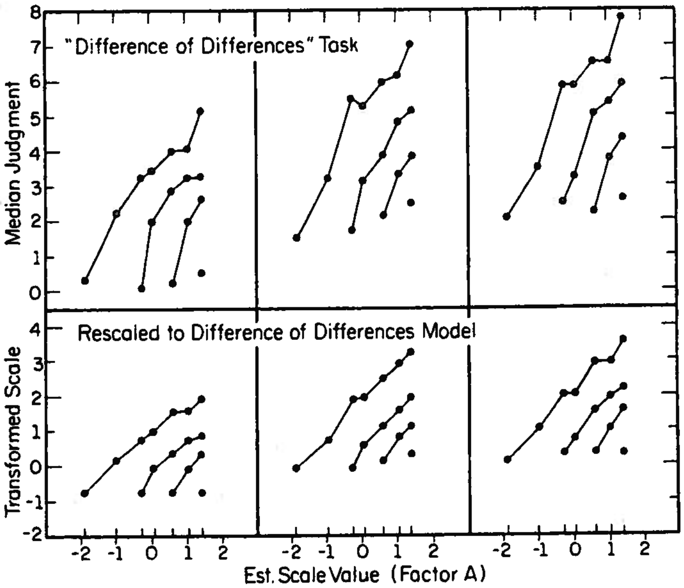

FIG. 16 Results of "difference of differences" task for likeableness judgments. Median ratings are plotted (only for positive differences) as a function of scale values for difference of differences model. Lower panel shows rescaled values, plotted in same fashion. (From Hagerty and Birnbaum, 1976.)

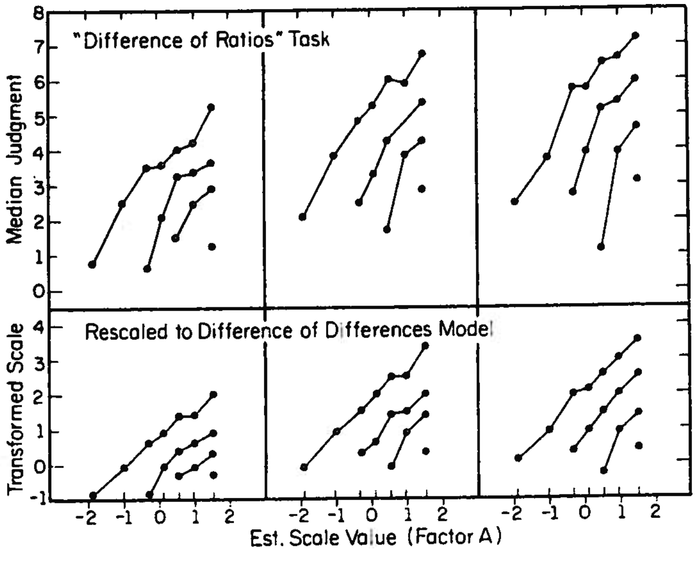

FIG. 17 Results of "difference of ratios" task. Data and rescaled values are plotted as in Fig. 16. Results are compatible with difference of differences model, not difference of ratios model. (From Hagerty and Birnbaum, 1976.)

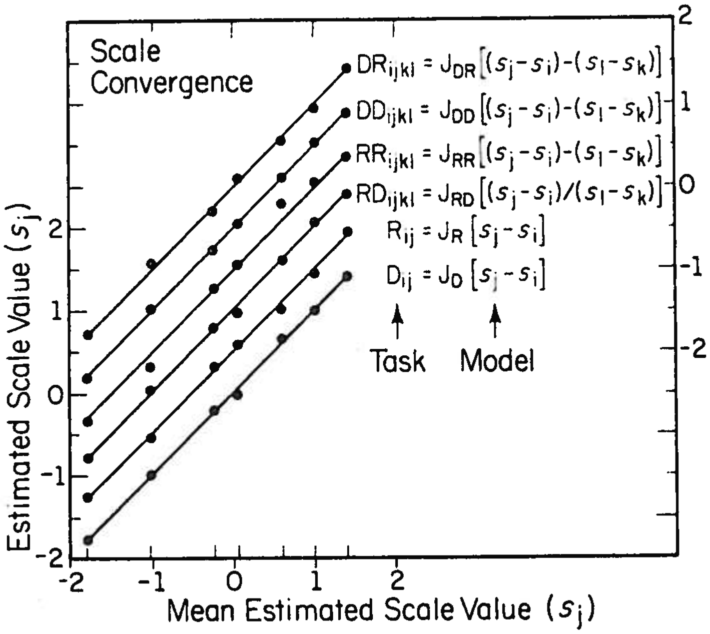

FIG. 18 Tests of scale convergence. Estimated scale values for likeableness of seven column adjectives as a function of mean estimated scale value. Each curve represents scale values derived from a set of data (task) using model shown to the right. Each curve has been displaced .5 units on the ordinate. Linear agreement of the scales is consistent with the theory that the models from which the scales are derived can be interlocked with the same scale values. (From Hagerty and Birnbaum, 1976.)

The most attractive description of all of the data is that the subtractive model applies to "ratios" and "differences" and that the difference of differences model represents not only "differences of differences" but also "ratios of ratios" and "differences of ratios." There is only one ratio operation in all of the models, for "ratios of differences," where the ratio of differences model is applicable. Figure 18 shows the scale values for the seven column adjectives estimated from these models, plotted as a function of the average of the scale value estimates. Each set of scale values has been shifted .5 units on the ordinate; identity lines have been drawn in to aid the examination of linearity.

The two lowest curves show that when the subtractive model is used to derive scales from simple "ratios" and "differences," the scale values are in close agreement with scales derived from the other tasks. The top two curves show that "differences of ratios" and "differences of differences" yield scales that are nearly linearly related to the others when fit to the difference of differences model. The third curve from the top shows that "ratio of ratios" judgments, even though they require drastic (approximate logarithmic) transformation, yield scales in approximate agreement with the others when the data are fit to the difference of differences model. If the subjects were truly computing ratios of ratios using common scale values, the plotted scale values (for the difference of difference model) would have been a logarithmic function of the other scale values. The "ratio of differences" task is the one that really specifies the system. Scale values derived from this model agree with the subtractive theory of all of the other tasks. To replace the subtractive operations in the models with ratios while retaining scale convergence would require that the ratio of differences model be replaced with the complex model of Eq. (13), in which there are two different scales and two different comparison processes for "differences," and in which "ratio" is modeled by a power equal to the reciprocal of a difference. Therefore, the subtractive theory (the set of models shown in Fig. 18) appears to give the simplest and most coherent account of all of the data in terms of a unified scale of likeableness for the adjectives.

TENTATIVE THEORETICAL CONCLUSIONS

The data reviewed here form a simple, consistent picture that justifies discussion of a rather different set of theoretical propositions from those of certain currently popular views. It would be helpful to see replications of the four-stimulus experiments such as those of Veit (1974, in press) and Hagerty and Birnbaum (1976) using other stimuli. If the results of these experiments hold up in further research, they go a long way toward explaining the long-standing controversy of "ratio" versus "interval" scales, clarifying the issues of stimulus comparison, stimulus representation, and judgment.

Comparison Processes

Torgerson's (1961) theory that subjects perceive only a single perceptual comparison between two stimuli was based in part on the approximately logarithmic relationship between category ratings and magnitude estimations (Torgerson 1960). It was also based on Garner's (1954) finding that subjects tended to make the same settings when instructed to adjust a tone to either "bisect" a loudness interval or to establish equal "ratios." Stronger evidence for the idea that there is only one operation is provided by results of factorial experiments in which different theories make different ordinal predictions for the data. In factorial experiments with loudness, darkness, likeableness, and heaviness judgments, it appears that judgments of "differences" and "ratios" of two stimuli are monotonically related, consistent with Torgerson's hypothesis that the comparison process is independent of the task (Birnbaum & Veit, 1974a; Birnbaum & Elmasian, 1977; Hagerty & Birnbaum, 1976; Rose & Birnbaum, 1975; Veit, 1974; see also Schneider et al., 1976).

Torgerson (1961) contended that if the subject appreciates only a single relationship between a pair of stimuli, it would not be possible to test empirically between distance and ratio interpretations of this relation. Some have concluded that it would never be meaningful to ask which representation was the "correct" one, since for a single two-factor design, ratio and subtractive models cannot be differentiated on the basis of ordinal tests. However, the scale-free tests possible with four-stimulus tasks, together with the criterion of scale convergence, provide the leverage to differentiate alternative theories of stimulus comparison. Instructions to judge "ratios" and "differences" *do* lead to two distinct judgment orders when the objects of judgment are stimulus differences. Scales values defined by the subtractive model for both two-stimulus tasks agree with those derived from the ratio of differences model applied to judgments of "ratios of differences" and they agree with scales derived from a difference of differences model applied to the other four-stimulus tasks. Since the results of the four-stimulus experiments interlock with the two-stimulus results, it appears that the process by which two stimuli are compared can best be represented by the subtractive model.

In summary, the following premises are consistent with the data:

P_1 (independence): The scale value of a stimulus is independent of the stimuli with which it is compared.

P_2 (scale convergence): The scale value of a stimulus is independent of tasks to judge "ratios," "differences," "ratios of ratios," "ratios of differences," "differences of ratios," or "differences of differences."

P_3 (magnitude estimation): $\mathrm{ME}_j = J_{\mathrm{M}}(s_j)$

P_4 (category judgments): $\mathrm{CJ}_j = J_{\mathrm{C}}(s_j)$

P_5: $R_{ij} = J_{\mathrm{R}}(s_j - s_i)$

P_6: $D_{ij} = J_{\mathrm{D}}(s_j - s_i)$

P_7: $DR_{ijkl} = J_{\mathrm{DR}}\,[(s_j - s_i) - (s_l - s_k)]$

P_8: $DD_{ijkl} = J_{\mathrm{DD}}\,[(s_j - s_i) - (s_l - s_k)]$

P_9: $RR_{ijkl} = J_{\mathrm{RR}}\,[(s_j - s_i) - (s_l - s_k)]$

P_{10}: $RD_{ijkl} = J_{\mathrm{RD}}\,[(s_j - s_i)/(s_l - s_k)]$

The first premise extends the independence assumption to the four-stimulus tasks. Premise P_2 extends the criterion of scale convergence to include all of the comparison tasks considered here; thus the values of s are assumed to be the same for all of the models. Premises P_3 and P_4 assert that magnitude estimations and category ratings of single stimuli are monotonic functions of subjective value. Premises P_5 and P_6 represent judgments of both "ratios" and "differences" of two stimuli with the subtractive model. Premises P_7, P_8, and P_9 represent the processes underlying three tasks, "differences of ratios," "differences of differences," and "ratios of ratios" with the difference of differences model. Premise 10 represents "ratios of differences" with a ratio of differences model. The judgment functions, J_{D}, J_{R}, J_{DR}, J_{DD}, etc., are assumed to be strictly monotonic.

One could ask the question, "Can a ratio theory be saved by replacing the operation of subtraction with division throughout?" The answer is that an exponential transformation of all of the models would yield a set of equations that would reproduce all of the data equally well. However, the "ratios of differences" task would be represented by the complex model of Eq. (13). This modified ratio theory seems too complicated to be seriously considered. Equation (13) not only violates the scale convergence criterion within itself, requiring two different scales, s and s^*, it suggests that two different models apply for "differences" within the same task. This theory represents "ratios" with either a ratio model (for "ratios") or an exponential-power model (for "ratios of differences"). Modified ratio theory also implies that the judgment functions for magnitude estimation are sometimes power functions (for "ratio" judgments) and sometimes logarithmic, since approximate parallelism in the left and right panels of Fig. 11 requires that $J_{\mathrm{R\,D}}$ be logarithmic for "ratios of differences." This theory seems as complicated as Tycho Brahe's geocentric theory of the solar system, which could give as good an account of the heavenly phenomena as Kepler's heliocentric theory if the laws of physics governing celestial events are allowed to be different from the laws describing earthly events. The argument that the earth revolves around the sun is based on simplicity, an assumed coherence between celestial mechanics and mechanics in the physics lab. In the same sense that Brahe's geocentric theory remains consistent with the data, so too does the complicated ratio theory.

It is helpful to show how another line of reasoning also leads to the subtractive theory. "Ratios" and "differences" of two stimuli can be expressed as $R_{ij} = J_{\mathrm{R}}\ (\Psi_{ij})$ and $D_{ij} = J_{\mathrm{D}}\ (\Psi_{ij})$, where the comparisons, Ψ_{ij}., are the same, but the monotonic judgment functions, J_{R} and J_{D}, are different. Ordinal analysis of the judgments indicates that the Ψ_{ij} form a group; consequently, one can write $\Psi_{ij} = s_j \circ s_i$, where $\circ$ is an unspecified operation. On the basis of two-stimulus judgments alone, Torgerson (1961) was correct in his assertion that the decision to represent $\circ$ with division or subtraction is "only a decision, not a discovery." However, results of four-stimulus experiments provide an empirical basis for testing between theories of the comparison operation. "Ratios of differences" and "differences of differences" demonstrate the appropriate ordinal requirements of ratio *and* difference operations on a common scale:

$$RD_{ijkl} = J_{\mathrm{RD}}\ [\Psi_{ij}/\Psi_{kl}] \tag{14}$$

$$DD_{ijkl} = J_{\mathrm{DD}}\ [\Psi_{ij} - \Psi_{kl}] \tag{15}$$

where the Ψ_{ij} values are the same in both equations and the judgment functions, J_{RD} and J_{DD}, are only assumed to be strictly monotonic. Without assuming anything about the comparison process, $\circ$, it is possible to use Eqs. (14) and (15) to derive values of Ψ_{ij}. These values will be unique to a ratio scale because they must reproduce both differences and ratios; hence the derived values cannot be subjected to nonlinear transformation. The Ψ_{ij} values thus derived are monotonically related to R_{ij} and D_{ij}, indicating that the same comparison operation, $\ominus$, can be used to represent both two- and four-stimulus judgments. The nature of this operation can be "discovered" by noting that the Ψ_{ij} values derived from Eqs. (14) and (15) are parallel (not bilinear) when plotted against the column values with a separate curve for each row. The parallelism implies that the operation by which two stimuli are combined is subtraction, $\Psi_{ij} = s_j - s_i$. It is conceivable that this plot could have been bilinear, which would have been consistent with ratio theory and inconsistent with subtractive theory. Thus the choice of the subtractive model is based on an empirical test and is not an a priori "decision."

Stimulus Representation

Since "ratios of differences" can be represented by a ratio of differences model, the failure of the ratio model for simple "ratio" judgments cannot be explained by asserting that subjects do not possess the "mental capacity" for two operations. Instead, the stimulus representation may be inherently no more than an interval scale, like points along a line in a subjective space (Veit, 1974). In this case, intervals are meaningful but ratios are not. For example, what is the ratio of the "easterliness" of New York to that of Denver? Without a well-defined zero point, the question does not make sense except in terms of distances. The following question does make sense: "What is the ratio of the distance from New York to Denver, relative to the distance from New York to San Francisco?" It may be that the subject thinks of degrees of darkness or likeableness in the same way that one thinks of locations on a map.[3] When instructed to judge "ratios," the subject cannot make sense of the task and reverts to computing differences. Only when there is a well-defined zero point, as in the case of "ratios of differences," does the subject actually compute ratios.

For certain continua, magnitude estimations and category ratings of single stimuli seem to agree. These continua were named "metathetic" to contrast them with the "prothetic" continua for which the two scales were nonlinearly related (Stevens & Galanter, 1957). A more fundamental distinction would be between continua for which "ratios" and "differences" generate only one or two distinct orderings, suggesting one or two comparison operations. Since stimulus intervals obey this criterion having two orders, one might expect that visual length, which can be thought of as a distance between points, might also allow two operations. Parker, Schneider, & Kanow (1975) represented "ratios" and

[3] In a recent experiment, done in collaboration with Barbara Mellers, subjects were indeed asked to make judgments of "ratios" and "differences" of easterliness and westerliness of U.S. cities. The results were consistent with the interpretation that the subtractive operation underlies all four tasks, with estimations of "ratios" exponentially related to subjective intervals.

"differences" of length with two operations, so length may indeed have a well-defined zero point. However, for loudness, likeableness, heaviness, and darkness, it appears that the stimulus representation may be inherently no more than an interval scale.

If subjects compute differences instead of ratios, why do the raw data for "ratios" fit a ratio model? The answer to this question requires a theory of the judgmental transformation.

Judgmental Processes

Premises P_1 through P_{10} account for the ordinal properties of the data. To account for the actual numerical judgments requires additional premises about the nature of the judgmental (J) functions. Although P_1 through P_{10} can be used to estimate these functions from the data, it seems useful to discuss potential explanations of judgment from which the J functions could be predicted.

Category ratings. Parducci's range-frequency theory has been successful in describing ratings of stimuli presented in varying stimulus distributions. (Parducci, 1974; Parducci & Perrett, 1971; Birnbaum 1974b). The theory assumes that ratings reflect a compromise between two tendencies: (1) judges tend to make differences in response proportional to differences in stimulus rank; and (2) judges tend to make differences in response proportional to differences in scale value.

The range-frequency model can be written (Birnbaum, 1974b):

$$\mathrm{CJ}_{jk} = (\mathrm{C}_\mathrm{m} - \mathrm{C}_o)\,[aG_k\,(s_j) + b(s_j - s_o)/(s_m - s_o)] + C_o, \qquad (16)$$

where CJ_{jk} is the category judgment of stimulus j in context k, on a scale from C_o to C_m; s_m and s_o are scale values of the maximum and minimum stimuli; $G_k\,(s_j)$ is the cumulative density of stimuli having scale values less than or equal to s_j in context k; and a and b are the weights of the frequency and range principles, respectively.

Range-frequency theory predicts that if the stimuli are spaced evenly on the subjective scale and presented with equal frequency, the response will be a linear function of subjective value. The judgmental transformations for the two- and four-stimulus ratings, J_D, J_DR, and J_DD, are all nearly linear, as evidenced by the near-parallelism in Figs. 6, 8, 13, 16, and 17. The J functions for category ratings have been found to be nearly linear in other studies involving subtractive models (Birnbaum, 1974a; Birnbaum & Veit, 1974a, 1974b).

Although the J functions estimated here are nearly linear, it seems reasonable to suppose that the stimulus distribution for stimulus pairs also affects the J function. Birnbaum, Parducci, and Gifford (1971, Experiment V) found evidence that the form of J in an information integration task can be manipulated in accord with range-frequency theory applied to the distribution of integrated impressions (Ψ). It is tempting to theorize that Eq. (16) would apply to ratings of "differences" with the substitution of $|\Psi|$ for s. Thus ratings of "differences" may be approximated by the equation:

$$D_{ij} = \delta_{ij}(\mathrm{D}_m - \mathrm{D}_o)\,[aG(|\Psi_{ij}|) + b\,|\Psi_{ij}|\,/\,\Psi_m] + \mathrm{D}_o, \qquad (17)$$

where D_{ij} is the rating of the subjective difference, $\Psi_{ij} = s_j - s_i$; D_m is the maximal response; D_o is the response for "no difference", $\delta_{ij} = -1$ if $\Psi_{ij} < 0$; $\delta_{ij} = 0$ if $\Psi_{ij} = 0$; $\delta_{ij} = 1$ if $\Psi_{ij} > 0$; Ψ_m is the maximum absolute difference in the experiment; and $G(|\Psi_{ij}|)$ is the cumulative density for the absolute difference.

Research is needed to establish the locus of contextual effects in information integration and to test the applicability of Eq. (17) in stimulus comparison experiments.

Magnitude estimation. To account for the approximate bilinearity of "ratio" judgments (Figs. 6, 7, and 13), the trilinearity of "ratios of ratios" judgments (Fig. 15), and the nonlinear relationship between ratings and magnitude estimations (Fig. 2), it is necessary to postulate that the judgmental transformations for magnitude estimation, J_{M}, J_{R}, and J_{RR}, are nearly exponential. An exponential transformation for J_{R} would cause a subtractive operation to lead to bilinear data, since if $R_{ij} = J_{\mathrm{R}}(s_j - s_i) = \exp(s_j - s_i)$, then $R_{ij} = \exp(s_j)/\exp(s_i) = s_j^*/s_i^*$, where $s^* = \exp(s)$.

Birnbaum and Veit (1974a) have proposed an interpretation of J_{R} that can account for an exponential transformation for magnitude estimation. The idea is shown in Fig. 19, which plots magnitude estimation responses against subjective

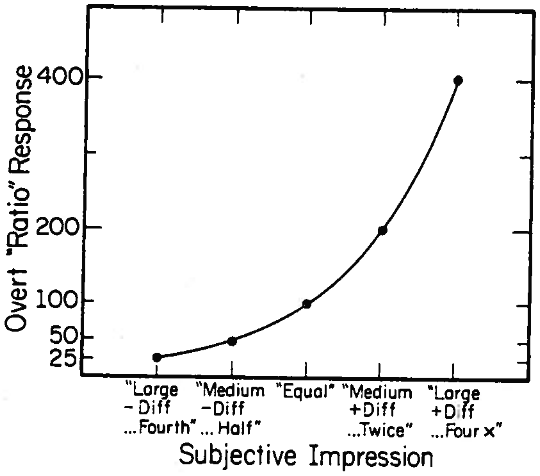

FIG. 19 Theory of the judgment function for magnitude estimations of "ratios." Abscissa represents the subjective continuum of comparisons (Ψ), evenly spaced with category labels of "ratios." It is assumed that reversing the stimulus order corresponds to equal *distances* from "equal" and that the distance from "equal" to "twice" is the same as the distance from "twice" to "four times." This process of response generation induces an exponential transformation. (After Birnbaum and Veit, 1974a).

differences varying from "large minus difference" through "zero difference" to "large positive difference." Suppose a subject is presented with a trial on which a "large difference" is presented. The subject selects a large magnitude estimation response, say "400." If the same pair of stimuli are presented in reverse order, the subjective difference would be the same but reversed in sign. However, the instructions require that the subject respond with the reciprocal "ratio," "25." If reversal in order corresponds to equal psychological distances and if the subject responds with reciprocal "ratios," then J_{R} will be positively accelerating as in Fig. 19. If, in addition, the subjective distance between "equal" and "twice" equals the subjective distance from "twice" to "four times," then the judgment transformation for magnitude estimation will be exactly exponential. Thus this theory of response generation explains how a ratio model could fit the data even though the comparison process is subtraction.

It is interesting to note that the largest mean "ratio" of loudness (Fig. 7), 5.69, is very nearly equal to the largest mean "ratio" of darkness (left of Fig. 6), 5.84. Teghtsoonian (1971) has discussed a theory of magnitude estimation in which the average log response range ($\log R_{\mathrm{max}} - \log R_{\mathrm{min}}$) is a constant. In spite of the judge's apparent freedom to choose any response range, Teghtsoonian notes that the average log response range is usually near 1.53. The log response ranges for darkness and loudness (Figs. 6 and 7) are about 1.65 and 1.43, respectively. Birnbaum and Elmasian (1977) found that subjects differed widely in the value assigned to the largest ratio. It may be useful to represent magnitude estimation in terms of an exponential transformation of Eq. (17), allowing the largest response, D_m, to depend on the subject. It may also be possible to manipulate the value of D_m through instructions, possibly in the examples given to illustrate the scale.

The judgmental transformations for magnitude estimations of "ratios of differences" have been estimated to be positively accelerated (Veit, 1974, Experiment II; Beck & Shaw, 1967) when the standard difference is intermediate in value. The curves in the upper-right panel of Fig. 14 show that on either side of zero, the J_{RD} function accelerates. This acceleration is consistent with a positively accelerated judgmental transformation for magnitude estimation. Yet, when the standard difference is the largest difference (Figs. 6 and 11 (C) and upper-left panel of Fig. 14), the J_{RD} function is nearly linear. Perhaps the J_{RD} function has a different form (a smaller slope) for response values less than 100 from the form it has for response values above 100.

A precise theory of magnitude estimation should predict how the subject selects his largest "ratio" response and how such factors as stimulus range, instructions, and modulus affect J. It may prove profitable to solve for J_{R} by rescaling "ratio" judgments to parallelism, under different conditions of context. For example, this procedure would allow a separation of the effects of stimulus spacing, on J and on s. To date, contextual effects are better understood for ratings than for magnitude estimations.

Summary

The data obtained from sets of factorial experiments suggest that the basic operation by which two stimuli are compared is subtraction. This conclusion depends on the premise that scales are independent of the judgmental task. The metric properties of the data satisfy the theory that magnitude estimations of "ratios" are an exponential function–and category ratings of "differences" are a linear function–of subjective differences. Consistent with the notion that the subjective stimulus representation is inherently an interval scale, "ratios of differences" can be represented by a ratio of differences model even though simple "ratios" are represented by subtraction.

ACKNOWLEDGMENT

This work was facilitated by support from the Research Board of the University of Illinois at Urbana-Champaign. Thanks are due to Robert Elmasian, Michael Hagerty, Barbara Mellers, Barbara Rose, Steven Stegner, and Clairice Veit, who have been collaborators in this research program and who have given helpful criticisms on earlier drafts.

REFERENCES

Anderson, N. H. Functional measurement and psychophysical judgment. *Psychological Review*, 1970, *77*, 153-170.

Anderson, N. H. Cross-task validation of functional measurement. *Perception and Psychophysics*, 1972, *12*, 389-395.

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

Attneave, F. Perception and related areas. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 4) New York: McGraw-Hill, 1962.

Beck, J., & Shaw, W. A. Ratio estimations of loudness intervals. *American Journal of Psychology*, 1967, *80*, 59-65.

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

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

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

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

Birnbaum, M. H., & Stegner, S. Ratios and differences of darkness. Unpublished experiment, University of Illinois, Urbana-Champaign, 1976.

Birnbaum, M. H., & Veit, C. T. Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception and Psychophysics*, 1974, *15*, 7-15. (a)

Birnbaum, M. H., & Veit, C. T. Scale-free tests of an additive model for the size-weight illusion. *Perception and Psychophysics*, 1974, *16*, 276-282. (b)

Cliff, N. Scaling. *Annual Review of Psychology*, 1973, *24*, 473-506.

Ekman, G. Two generalized ratio scaling methods. *The Journal of Psychology*, 1958, *45*, 287-295.

Ekman, G., & Sjoberg, L. Scaling. *Annual Review of Psychology*, 1965, *16*, 451-474.

Garner, W. R. A technique and a scale for loudness measurement. *Journal of the Acoustical Society of America*, 1954, *26*, 73-88.

Garner, W. R., Hake, H. W., & Eriksen, C. W. Operationism and the concept of perception. *Psychological Review*, 1956, *63*, 149-159.

Hagerty, M. & Birnbaum, M. H. Nonmetric tests of ratio vs. subtractive theories of stimulus comparison. Unpublished experiment, University of Illinois, Urbana-Champaign, 1976.

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

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. *Foundations of measurement*. New York: Academic Press, 1971.

Krantz, D. H., & Tversky, A. Conjoint measurement analysis of composition rules in psychology. *Psychological Review*, 1971, *78*, 151-169.

Kruskal, J. B. & Carmone, F. J. MONANOVA: A FORTRAN-IV program for monotone analysis of variance. *Behavioral Science*, 1969, *14*, 165-166.

Marks, L. E. On scales of sensation: Prolegomena to any future psychophysics that will be able to come forth as a science. *Perception and Psychophysics*, 1974, *16*, 358-376.

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

Parker, S , Schneider, B., & Kanow, G. Ratio scale measurement of the perceived lengths of lines. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *104*, 195-204.

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

Rose, B. J., & Birnbaum, M. H. Judgments of differences and ratios of numerals. *Perception and Psychophysics*, 1975, *18*, 194-200.

Rule, S. J., & Curtis, D. W. Conjoint scaling of subjective number and weight. *Journal of Experimental Psychology*, 1973, *97*, 305-309.

Sarris, V., & Heineken, E. An experimental test of two mathematical models applied to the size-weight illusion. *Journal of Experimental Psychology: Perception and Performance*, 1976, *2*, 295-298.

Savage, C. W. Introspectionist and behaviorist interpretations of ratio scales of perceptual magnitudes. *Psychological Monographs*, 1966, *80*, 1-32.

Schneider, B., Parker. S., Kanow, G., & Farrell, G. The perceptual basis of loudness ratio judgments. *Perception and Psychophysics*, 1976, *19*, 309-320.

Shepard. R. N. On the status of "direct" psychological measurement. In Savage (Ed.), *Minnesota Studies in the Philosophy of Science* (Vol. IX). Minneapolis: University of Minnesota Press, 1976.

Sjoberg, L. Sensation scales in the size-weight illusion. *Scandinavian Journal of Psychology*, 1969, *10*, 109-112.

Stevens, J. C., & Rubin, L. L. Psychophysical scales of apparent heaviness and the size-weight illusion. *Perception and Psychophysics*, 1970, *8*, 225-230.

Stevens S. S. On the psychophysical law. *Psychological Review,* 1957, *64,* 153-181.

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,* 377-411.

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

Torgerson, W. S. Quantitative judgment scales. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications.* New York: Wiley, 1960.

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

Treisman, M. Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology,* 1964, *16,* 11-22.

Veit, C. T. *Ratio and subtractive processes in psychophysical judgment.* Unpublished doctoral dissertation, University of California, Los Angeles, 1974.

Veit, C. T. Ratio and subtractive processes in psychophysical judgment. *Journal of Experimental Psychology: General,* in press.

Weiss, D. J. Averaging: An empirical validity criterion for magnitude estimation. *Perception and Psychophysics,* 1972, *12,* 385-388.