Perception & Psychophysics
1978, Vol. 24 (2), 121-129

Nonmetric tests of ratio vs. subtractive theories of stimulus comparison

MICHAEL HAGERTY and MICHAEL H. BIRNBAUM
University of Illinois, Champaign-Urbana, Champaign, Illinois 61820

Theories of stimulus comparison were tested by examining ordinal properties of data obtained with six scaling tasks. Subjects judged simple "ratios" or "differences" of stimulus pairs constructed from a factorial design. In four additional tasks, the same judges also compared relations between pairs of stimulus pairs, judging "ratios of ratios," "ratios of differences," "differences of ratios," and "differences of differences." The data were consistent with a subtractive theory, which asserts that two stimuli are compared by subtraction, regardless of the task, but that judges can compare two stimulus *differences* by either a ratio or a difference. All six tasks could be related by the subtractive theory using a single set of scale values. Other simple theories, including the theory that "ratio" judgments can be represented by a ratio model, could not reproduce the six rank orders (of the six sets of data) using a single set of scale values.

A long-standing puzzle in the history of psychophysics and scaling has been the lack of agreement between category judgments and magnitude estimations—between so-called "ratio" and "interval" techniques for scaling.[1] This paper offers a theoretical and empirical structure to resolve this puzzle.

Judgment Functions

Let CJ_i and ME_i be the category judgment and magnitude estimate of stimulus i, having subjective scale value, s_i. We assume that the two dependent variables are at least monotonically related to psychological values:

$$CJ_i = J_C(s_i) \quad (1)$$

$$ME_i = J_M(s_i), \quad (2)$$

where J_C and J_M are strictly monotonic judgment functions for the category rating task and the magnitude estimation task. Equations 1 and 2 clarify the distinction between the subjective value of a stimulus and the numerical response given to it by the subject. If the number of categories is changed from 9 to 20, for example, one would not want to suppose that the sensation changed. Although the rank order of responses is assumed to be the same as the rank order of scale values, it is not assumed that responses are a linear function of subjective values.

The relationship between ratings and magnitude estimations is the composition:

$$CJ_i = J_C[J_{\bar{M}}^{-1}(ME_i)]. \quad (3)$$

If the judgment functions were both linear, J_C and $J_{\bar{M}}^{-1}$ would be linear and category ratings would be a linear function of magnitude estimations. Instead, it appears that this composition of functions, $J_C J_{\bar{M}}^{-1}$, is concave downwards (Stevens & Galanter, 1957). Torgerson (1960) observed that this function is often approximately logarithmic.

Torgerson's Theory: One Operation

Torgerson (1961) suggested that the contradiction between scaling results occurs not just because the judges have two contrary scales, but also because the experimenters have two different tasks and believe that their subjects have two corresponding operations. Torgerson pointed out that if judges have only one way of comparing two stimuli, the experimenter who thinks the relation is a ratio will derive a scale of sensation that is exponentially related to the scale derived by the experimenter who represents the relation as a difference. Two experimenters could start with the same data and derive two different scales.

Unfortunately, Torgerson's (1961) suggestion was criticized on the basis of a red herring: category judgments and magnitude estimations (two so-called "direct" measures) are not exactly exponentially related in every experiment. However, it is reasonable to theorize that J_C and J_M depend upon the stimulus spacing, range, relative frequency of presentation, and other contextual details of the experiment (Birnbaum, 1974b, 1978; Parducci, 1963, 1974; Parducci & Perrett, 1971; Poulton, 1968). Therefore, the relationship between category ratings and magnitude estimates (the composition, $J_C J_{\bar{M}}^{-1}$) is expected to vary from situation to situation.

These experiments were supported by a grant from Research Board, University of Illinois. Requests for reprints should be addressed to Michael H. Birnbaum, Department of Psychology, University of Illinois, Champaign, Illinois 61820.

Fortunately, Torgerson's theory can be tested without having to assume metric properties in the response scale. In a factorial design, actual ratios and differences are *not* monotonically related (e.g., $7/5 < 2/1$ but $7-5 > 2-1$). Therefore, if subjects can distinguish instructions to judge "ratios" and "differences," the two types of judgments should *not* be monotonically related, in general, but instead the two distinct rank orders should be appropriately interrelated (Krantz, Luce, Suppes, & Tversky, 1971, p. 152-154). On the other hand, if there is only one operation for both tasks, then both types of judgments will be monotonically related.

Experiments: One Operation

Recent research with a variety of dimensions shows that magnitude estimates of "ratios" and category ratings of "differences" are monotonically related (Birnbaum, 1978, Birnbaum & Elmasian, 1977; Birnbaum & Mellers, 1978; Birnbaum & Veit, 1974; Rose & Birnbaum, 1975; Veit, 1978). These results are consistent with the theory that judges use only one operation to compare stimuli, regardless of instructions to judge "ratios" or "differences" and regardless of the procedure for responding. Since actual ratios and differences of the same numbers are *not* monotonically related in suitable factorial designs, these experiments provide nontrivial support for Torgerson's (1961) theory that both tasks are governed by the same comparison operation.

Indeterminacy?

Torgerson (1961) noted that if judges do perceive only one type of stimulus relationship, it would not be possible to determine its nature in a single, two-factor experiment. To represent this single comparison operation as a difference or ratio would be a "decision, not a discovery." However, Birnbaum and Veit (1974; Birnbaum, 1978; Veit, 1978) have shown that ratio and subtractive theories make distinct predictions in conjunction with a few extra, but reasonable, constraints involving a wider array of data.

Scale Convergence Criterion

One such constraint is the principle of scale convergence (Birnbaum, 1974a; Birnbaum & Veit, 1974). Scale convergence is the premise that the scale values of the same stimuli derived from models of different empirical relationships should agree. A theoretical structure of two or more empirical structures satisfying scale convergence for the same stimuli is preferable to an alternative theory that violates scale convergence.

Rose and Birnbaum (1975) showed that scale values for numerals derived from the ratio model are a positively accelerated function of physical number. However, scale values for the subtractive model applied to the same data were a negatively accelerated function of number and were in close agreement with a scale for the same stimuli derived from Parducci's range-frequency theory (Birnbaum, 1974b) and with scales derived by others (Rule & Curtis, 1973). Thus, the subtractive model is preferred, since it yields scales that are consistent with a family of other theories, whereas the ratio model violates scale convergence.

Birnbaum and Veit Theory: Subtraction

Birnbaum (1978) and Veit (1978) have proposed that for many continua, stimuli should be represented as positions on an interval scale rather than magnitudes on a ratio scale. If so, ratios are not meaningful. Birnbaum and Veit proposed that when the stimuli do not have a well-defined zero point, judges will compute differences whether instructed to judge "differences" or "ratios." For example, the question, "What is the ratio of the easterliness of Philadelphia relative to that of Denver?" does not make sense unless a zero point is arbitrarily located on the mental map. Birnbaum and Mellers (1978) actually asked judges to make such (seemingly) strange judgments. The results, interestingly enough, were similar to judgments of "ratios" of loudness: The ratio model gave a good fit to the data. However, the data were *not* consistent with the hypothesis that judges insert zero points for judging "ratios" of "easterliness" and "westerliness." Instead, mental maps based on the ratio model applied to "ratio" judgments were exponentially distorted, depending on the direction of judgment. A subtractive model, though it required approximately logarithmic transformation of the numerical judgments, gave a good account of the data. The subtractive model yielded a single mental map (which resembled the actual map) that was independent of the direction of judgment.

Scale-Free Tests

Even though judges might not be able to compute ratios of sensations such as loudness of tones or easterliness of cities, perhaps they can judge ratios of stimulus intervals. For example, even though the question, "What is the ratio of the easterliness of Philadelphia to that of Denver?" may not have a meaningful answer, the following question *is* meaningful on an interval scale: "What is the ratio of the distance from Denver to Philadelphia relative to the distance from San Francisco to Philadelphia?" An interval scale permits both ratios and differences of intervals. Veit (1978) used the "ratio of differences" task to provide a scale-free test between ratio and subtractive theories of stimulus comparison. Her data were consistent with a ratio of differences model, and favored the subtractive theory.

The present research extends Veit's (1978) approach to testing differential predictions of subtractive and ratio theories. By simultaneously studying a set of tasks employing factorial designs of four stimuli, it is possible to test several possibilities left untested in Veit's (1978) research. The subjects are instructed to compare the relation between one pair of stimuli relative to another pair relation. There are four such tasks: (1) "*ratio* of two *differences*," $(A-B)/(C-D)$; (2) "*ratio* of two *ratios*," $(A/B)/(C/D)$; (3) "*difference* between two *ratios*," $(A/B)-(C/D)$; and (4) "*difference* between two *differences*," $(A-B)-(C-D)$. Birnbaum (1978) has shown that these four-stimulus polynomials, which can be distinguished by nonmetric analyses (Krantz et al., 1971; Krantz & Tversky, 1971), provide the possibility of testing among different theories of stimulus comparison. The present experiment also includes the simple, two-stimulus "ratio" and "difference" tasks.

By the criterion of scale convergence, scale values derived from one set of data should agree with scale values for the same stimuli derived from another. Disagreement of scales permits rejection of a *set* of models, even if local tests are satisfied.

Potential Outcomes

There are a number of potential outcomes, each of which would be consistent with a different theory of stimulus comparison (Birnbaum, 1978). Three of the simpler possibilities follow.

First, it may be that subjects compare two stimuli by *subtraction*, and can compare two intervals by either a ratio or a difference. Under this subtractive theory, "ratios of differences" and "differences of differences" will produce two distinct orderings, consistent with the task-defined models. The resulting scale of *intervals*, which is defined to a ratio scale, should agree with the subtractive interpretation of the two-stimulus tasks.

Second, it may be that two stimuli are compared as a *ratio* and that once the ratio is computed, both difference and ratio operations are possible. If so, then the "difference of ratios" and "ratio of ratios" tasks should possess two rank orders which define a scale of ratios. These ratios, in turn, should agree with the ratio interpretation of the simple "ratio" and "difference" judgments.

Third, it may be that judges can only compare *two stimuli or two relations* by one operation. If so, then all of the four-stimulus tasks should be monotonically related, and all six tasks can be represented by one (indeterminate) operation.

Subtractive Theory of Six Tasks

The subtractive theory of stimulus comparison (Birnbaum, 1978; Birnbaum & Veit, 1974; Veit, 1978) can be made more explicit:

P_1: The scale value of a stimulus is independent of the task and the procedure for responding (scale convergence),

P_2: $R_{ij} = J_R(s_j - s_i)$,

P_3: $D_{ij} = J_D(s_j - s_i)$,

P_4: $RR_{ijkl} = J_{RR}[(s_j - s_i) - (s_l - s_k)]$,

P_5: $DR_{ijkl} = J_{DR}[(s_j - s_i) - (s_l - s_k)]$,

P_6: $DD_{ijkl} = J_{DD}[(s_j - s_i) - (s_l - s_k)]$, and

P_7: $RD_{ijkl} = J_{RD}[(s_j - s_i)/(s_l - s_k)]$,

where R_{ij} and D_{ij} are judgments of the "ratio" of stimulus level j to i and the "difference" between stimuli j and i; and RD, RR, DD, and DR are judgments of "ratios of differences," "ratios of ratios," "differences of differences," and "differences of ratios," respectively. The J functions represent monotone judgmental transformations relating overt judgments to subjective impressions.

The first premise makes explicit the scale convergence criterion; i.e., that the values of s_j should be identical in all of the tasks. Premises 2 and 3 assert that "ratios" and "differences" are both produced by a subtractive comparison process. Premises 4, 5, and 6 assert that whether instructed to judge "ratios of ratios," "differences of ratios," or "differences of differences," a subject will compare subjective intervals by subtraction. Note that Premises 4 and 5 follow from the subtractive theory that the subject will compare *two stimuli* by subtraction when the task is to compute a "ratio." Premise 7 asserts that judges use a ratio operation to compare *intervals* when instructed to judge "ratios of differences." Premises 6 and 7 follow from the subtractive theory, which postulates that judges can compute both ratios and differences of stimulus intervals.

METHOD

Participants first attended a 1-h session in which they were trained in four tasks, "ratios" (R), "differences" (D), "ratio of ratios" (RR), and "differences of differences" (DD). Training trials for four tasks were checked for reliability, consistency for special cases, and proper use of response scales.[2] About half the judges practiced D first, followed by R, DD, and RR. The others were trained in the sequence R, D, RR, and DD. The judges returned for two additional 2-h sessions to perform all six tasks, including "ratios of differences" (RD) and "differences of ratios" (DR).

Tasks

D: Judge the difference in likeableness between two separate people, described by the two adjectives. The 9-point scale had labels varying from 9 = "Like the person on the left very, very much more than the person on the right," to 1 = "Like the

person on the right very, very much more than the person on the left," with 5 = "Like both equally."

R: Judge the ratio of the likeableness of the first person to that of the second. The modulus was 100. Seven examples of geometrically spaced "ratios" were given ranging from 12.5 (first person is 1/8 as likeable as the second) to 800 (first person is 8 times as likeable as the second).

DD: Judge the *absolute* difference in likeableness between the two left-hand stimuli, the *absolute* difference in likeableness between the two right-hand stimuli, and rate the difference between the two differences. The 9-point scale had labels varying from 9 = "Left difference is very, very much greater than the right difference," to 1 = "Left difference is very, very much smaller than the right difference," with 5 = "Differences are equal."

RR: Judge the ratio of likeableness for the pair on the left, the ratio of likeableness for the pair on the right, and record the ratio of the first ratio to the second. The modulus was 100. Examples of seven geometrically spaced responses were given ranging from 12.5 (first ratio is 1/8 as great as second) to 800 (first ratio is 8 times greater than second ratio).

RD: Judge the difference in likeableness for the left-hand pair, judge the difference in likeableness for the right-hand pair, and judge the ratio of the left difference to the right, responding with a modulus of 100. Examples of ratio responses were given ranging from 0 (left-hand difference is zero) to ±800 (difference on left is 8 times greater than difference on right).

DR: Half of the subjects were instructed to judge the ratio of the *more* likeable to the *less* likeable person (for the pair on the left), judge the ratio of likeableness for the pair on the right, and judge the difference between the two ratios by using a 9-point scale, where 9 = "Left ratio is very, very much greater than the right ratio," 1 = "Left ratio is very, very much smaller than the right ratio," and 5 = "Ratios are equal."

The other half of the judges in the DR task (12) were given instructions to judge the ratio of the first stimulus to the second and subtract the ratio of the third stimulus to the fourth (comparable to RR and RD tasks) instead of the instructions above (which are comparable to the DD task). This instruction should theoretically not affect the data for trials on which the first stimulus exceeds the second. However, it should change the order of pairs below the diagonal, where the first stimulus is less. Separate analyses were performed for this portion of the design for these two groups.

In the DD, RD, and DR tasks, judges also indicated which adjective of the first pair was preferred by using a plus or minus sign if they preferred the first or second adjective, respectively.

Designs

The two two-stimulus tasks, D and R, used 7 by 4, First Adjective by Second Adjective, factorial designs. First adjectives included: *cruel*, *irritating*, *clumsy*, *hesitant*, *thrifty*, *capable*, and *sincere*. Second adjectives included: *mean*, *untidy*, *excited*, and *honest*. The adjectives were chosen from Anderson's (1968) list to be equally spaced on ratings of likeableness.

For the four four-stimulus tasks, the 4 by 7 design was combined factorially with three subtrahend (or divisor) pairs (*truthful – phony*, *truthful – listless*, and *practical – listless*), yielding a (4 by 7) by (3) design, with a total of 84 trials for each four-stimulus task. The subtrahend (divisor) pairs were chosen so that the first adjective would be more likeable in normative value.

Procedure

One replicate of the entire experiment was printed in a booklet, including the two two-stimulus tasks and the four four-stimulus tasks, with instructions for each task preceding the stimuli. Half the booklets contained tasks in the sequence D, R, DD, RR, DR, and RD. The remaining booklets contained tasks in the sequence R, D, RR, DD, RD, and DR. Further, three different random sequences of stimuli were produced by permuting the pages within each task, resulting in a 2 by 3 factorial of task sequence by stimulus order.

Each pair of adjectives appearing on the left-hand side of the four-stimulus tasks (as a dividend or minuend pair) was judged a total of 14 times: 3 times (in conjunction with three different divisors or subtrahends) in each of the four four-stimulus tasks, and once in each of the two two-stimulus tasks, giving $3 \times 4 + 2 = 14$ appearances. In 11 of these appearances, responses indicate which adjective of the left-hand pair was more likeable. For the DD, RD, and DR tasks, in which the subjects used minus signs to make this indication, 6% of the signs were reflected to make them consistent with the subject's majority sign. Median responses were then computed (using linear interpolation) for each cell in each design.

Subjects

Twenty-three undergraduates of University of Illinois at Urbana-Champaign participated for three sessions totalling 5 h. They received extra credit in a psychology course. Fifteen of these subjects were able to complete two repetitions of the entire design. The remaining eight did not complete the entire second repetition, so only their first replicate data were retained.

RESULTS

"Differences" and "Ratios"

Medians for the "ratio" and "difference" tasks are shown in Figure 1, plotted as a function of marginal means for the first adjective. Separate curves are used for different levels of the second adjective. In the left panel, "ratio" estimations show the approximate bilinear form predicted by the ratio model. The right panel shows that the "difference" ratings are approximately parallel, as predicted by the subtractive model.

The "ratio" medians and the "difference" medians were separately rescaled by MONANOVA, a computer program for monotone transformation (Kruskal & Carmone, 1969) which maximizes the fit to the additive (in this case, subtractive) model.[3] The transformed medians of "differences" (points) and "ratios" (circles) are plotted in the center panel

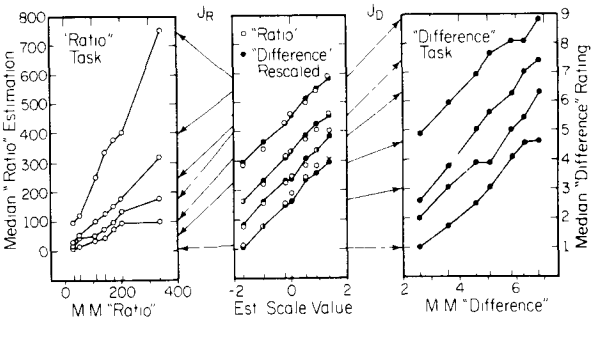

Figure 1. "Ratios" and "differences" in likeableness of adjectives. Left-hand panel shows median estimation of "ratio," plotted against marginal values for the first adjective with a separate curve for each level of the second. Right-hand panel shows median rating of "difference" plotted in the same fashion. Center panel 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 transformations.

as a function of the average of the marginal means of the transformed values. Both sets of transformed medians are approximately parallel, consistent with either a subtractive or a ratio model. Further, the two sets of transformed scores are nearly identical, consistent with the interpretation that one comparison operation underlies both tasks.

"Ratio of Differences"

The upper panels of Figure 2 present the median judgments for the "ratio of differences" task. The responses are plotted as a function of the marginal means for the seven adjectives of the first factor. Data for the largest divisor difference (*truthful-phony*) are on the left. The curves show the form predicted by the ratio of differences model: the smaller the divisor difference, the greater the slopes of the curves and vertical spreads between the curves.

A separate 4 by 7 MONANOVA rescaling was performed for each divisor difference. The rescaled medians are shown in the lower panels of Figure 2. The parallelism indicates that each numerator can be represented as a difference.

If the difference of differences or the ratio of ratios models could represent these data, then it should be possible to rescale the entire 7 by 4 by 3 design to parallelism. Instead, MONANOVA could not rescale the data for the entire design (the plotted rescaled values were not parallel).

The reason for MONANOVA's failure to rescale

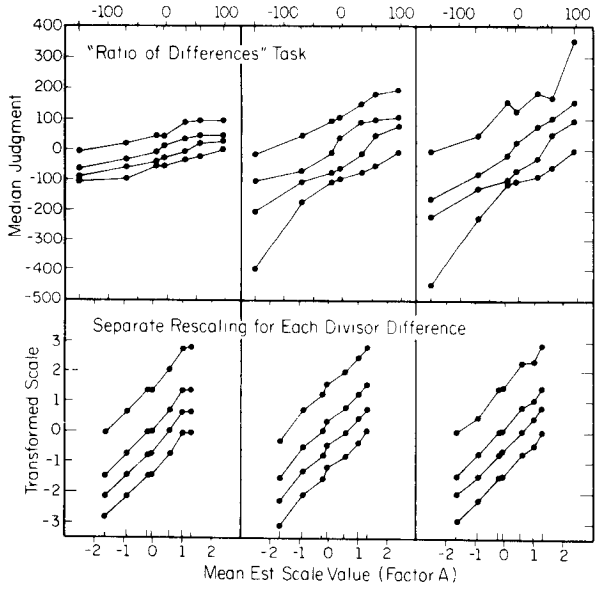

Figure 2. "Ratios of differences" in likeableness. Upper panels plot median estimates, with a separate panel for each denominator difference. Largest divisor difference is on the left, smallest on the right. Lower panels plot rescaled values with a separate rescaling for each panel. Data are compatible with a ratio of differences model.

the data for "ratio of differences" (Figure 2) can be explained as follows. The ratio of ratios and difference of differences models are ordinally additive in form and therefore predict that the rank order for each pair of factors will be jointly independent of the level of the third (Krantz & Tversky, 1971). The ratio of differences model, however, is a distributive model $[(A-B)/(C-D)]$, in which the rank order of the first stimulus (A) by the denominator $(C-D)$ matrix should depend on the level of the second stimulus (B). Violations of joint independence can be seen by plotting (or imagining) the data for all three upper panels of Figure 2 in the same panel. The ratio of differences model correctly predicts that the sets of curves should cross, and that the crossovers should occur when the first stimulus equals the second $(A-B=0)$. These crossover interactions cannot be rescaled to parallelism by any monotonic transformation. Therefore, the ratio of ratios and difference of differences models cannot account for the "ratio of differences" data. However, the ratio of differences model remains consistent with the data.

A final check on the model diagnosis tested the form of the Dividend by Divisor interaction. Assuming that the ratio of differences model is appropriate, then the theoretical value of each of the 28 dividends can be estimated by averaging across the three rescalings in Figure 2. Furthermore, the 28 by 3, Dividend by Divisor, interaction should be multiplicative. Therefore, if this bilinear dividend/divisor interaction is rescaled to parallelism, the rescaled positive numerators should be a *negatively accelerated* function (log) of the average difference (from the lower panels of Figure 2). If the difference of differences or ratio of ratios models held, however, the rescaled numerators from a numerator by denominator rescaling should be a *linear* function of the differences. Finally, if the difference of ratios model were appropriate, the rescaled dividends would be a *positively accelerated* function of the numerator differences. To check these mutually exclusive predictions, the positive portion of the 28 by 3 Dividend by Divisor interaction was rescaled by MONANOVA. These rescaled dividends were a negatively accelerated function of the numerator differences from Figure 2 consistent with the ratio of differences model.

In summary, the "ratio of differences" data are consistent with the ratio of differences model (and inconsistent with the other three simple polynomials), in agreement with the subtractive theory and with the results of Veit (1978).

"Ratio of Ratios"

Results for the "ratio of ratios" task are shown in Figure 3, plotted as in Figure 2. The median judg-

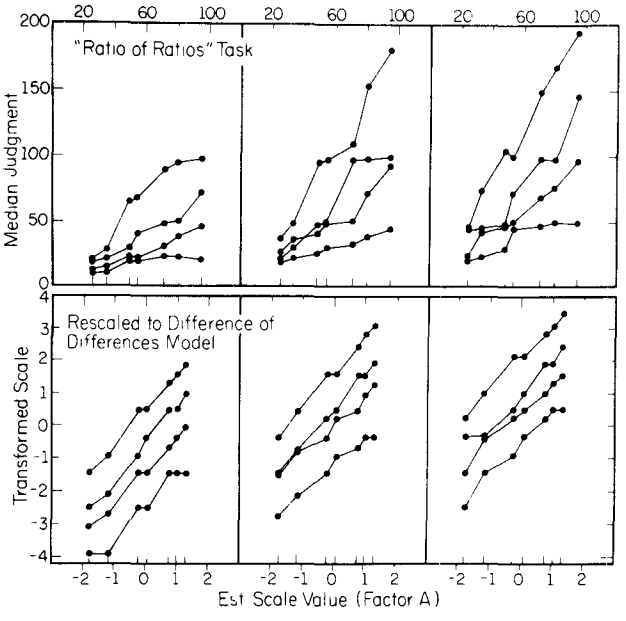

Figure 3. Results of "ratio of ratios" task, plotted as in Figure 2. Lower panels plot rescaled medians, fit to the difference of differences model. Scale values for the difference of differences model agree with scale values for the ratio of differences model fit in Figure 2.

ments, plotted in the upper panel as a function of the marginal means of the first adjective, approximate the trilinear divergent interaction predicted by the ratio of ratios model. The lower panels show the medians, rescaled by MONANOVA to fit the difference of differences model. Parallelism, linearity, and congruence of the three sets of curves would constitute 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.

"Difference of Differences"

Figure 4 shows median ratings of "difference of differences" for the positive portion of the design [that is, that portion of the 7 by 4 design (16 out of 28 cells) in which the first adjective was rated more likeable than the second adjective more than half of the time]. The rescaled medians are plotted as a function of marginal means of the first adjective in the lower panels. The near linearity, parallelism, and congruence of the sets of curves is consistent with the predictions of the difference of differences model. The negative portion of the design appears much as the positive does, and can also be rescaled to parallelism.

"Difference of Ratios"

Median ratings of "difference of ratios" are shown in Figure 5, plotted for the positive portion as in Figure 4. The data do not conform to the predictions of the difference of ratios model, which predicts diverging fans for each set of curves (Birnbaum, 1978). Rescaling A/B − C/D as a 16 by 3 to fit the subtractive model should result in divergent A/B fans, but instead, the data yielded approximately parallel curves. Indeed, the curves appear very similar to the data for the "difference of differences" task above and can be rescaled to fit the difference of differences model by a nearly linear transformation. This rescaling yields transformed values (lower panels of Figure 5) that are nearly

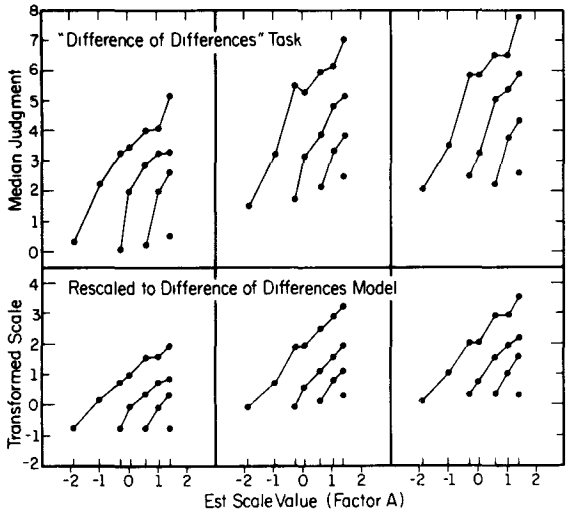

Figure 4. Results of "difference of differences" task. Median ratings are plotted (only for positive differences) as a function of scale values for the difference of differences model. Lower panel shows rescaled values, plotted in same fashion.

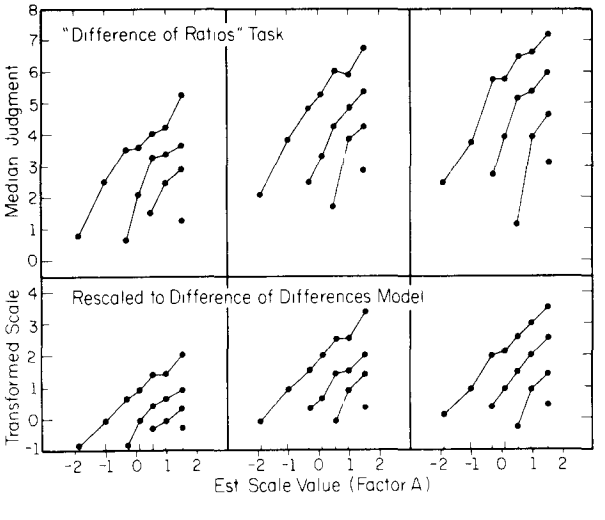

Figure 5. Results of "difference of ratios" task. Data and rescaled medians are plotted as in Figure 4. Results are compatible with the difference of differences model, not with the difference of ratios model.

congruent with transformed "difference of differences" data in Figure 4, consistent with the interpretation that the judges do not distinguish the two tasks. Separate analyses on the lower portion of the design for the two subgroups of judges who received different instructions also indicated that the "difference of ratios" judgments were like differences of differences for both groups.

Scale Convergence Tests

The theory that "ratio" and "difference" judgments in Figure 1 can be taken at face value is inconsistent with the scale convergence criterion. The scale values for the ratio model applied to the "ratio" judgments are .31, .49, .88, 1.06, 1.40, 1.87, and 2.67 for the seven column stimuli and .36, .85, 1.45, and 2.26 for the four row stimuli. These values, based on the assumption that J is a similarity transformation, give good predictions of the numerical "ratio" judgments. For example, the ratio, 2.67/.36 = 7.42, is very close to the largest obtained "ratio" judgment in Figure 1, 7.50. However, differences between these scale values do not reproduce the rank order of the "difference" judgments. For example, the ratio model scale values predict that the difference in likeableness between *sincere* and *excited* (2.67 − 1.45 = 1.22) should be greater than the difference between *thrifty* and *mean* (1.40 − .36 = 1.04). The "difference" ratings in Figure 1 show the opposite: most judges rate the "difference" in likeableness between *thrifty* and *mean* to be the greater.

The ratio model scale values also fail to reproduce the rank order of the data in Figure 5 when differences between ratios are calculated. Assuming that the subtrahend ratios are 7.42, 3.13, and 2.20, the ratio model scale values predict that all of the differences of ratios in the rightmost panel of Figure 5 should exceed all but the largest point in the leftmost panel. However, 7 out of the 16 points in the panel violate this prediction.

The subtractive theory, Premises P1-P7, gives a coherent account of the data. Figure 6 provides a summary of tests of scale convergence for the subtractive theory. Figure 6 shows the scale values for the seven column adjectives estimated from these models, plotted as a function of the average of the scale values estimates, −1.80, −1.02, −.24, .04, .62, 1.0, and 1.40. Each set of scale values has been vertically shifted .5 units on the ordinate; identity lines have been drawn to aid the examination of linearity.

The two lowest curves in Figure 6 show that when the subtractive model is used to derive scales from simple "ratios" and "differences," the estimated scale values are in close agreement with scales derived from the other tasks. If subjects were truly judging *both* ratios and differences of the same scale values,

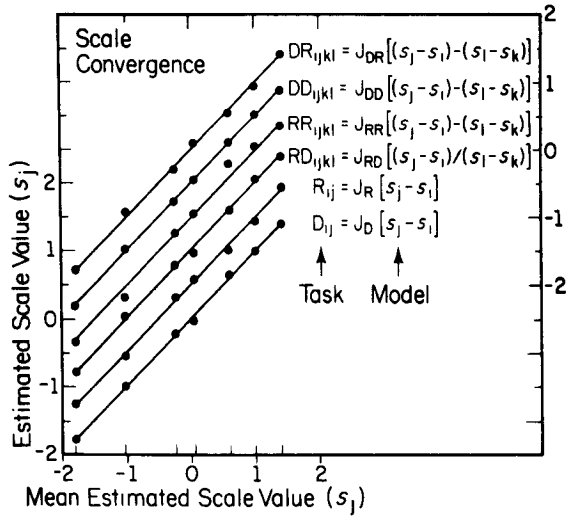

Figure 6. Tests of scale convergence for the subtractive theory. Estimated scale values of likeableness of seven column adjectives are plotted as a function of mean estimated scale value. Each curve represents scale values derived from a set of data (task) using the model shown to the right. Curves have been vertically 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.

then the two sets of scale value estimates would be expected to be logarithmically, *not* linearly, related.

The top two curves in Figure 6 show that "differences of ratios" and "differences of differences" yield scales that are linearly related to the others when fit to the difference of differences model. If subjects were truly performing two different operations, these two sets of scale value estimates would be nonlinearly related.

The third curve from the top shows that "ratio of ratios" judgments, even though they require drastic rescaling (approximately logarithmic) are in fair 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 differences model) would have been a logarithmic function of the other scale values.

The "ratio of differences" task specifies the entire system, since it contains both a ratio and subtractive operation for intervals and hence cannot be rescaled to another simple polynomial. Scale values derived from this model agree with the subtractive theory of all of the other tasks.

In sum, the near linearity of the curves in Figure 6 indicates that the subtractive theory is consistent with the six sets of data in terms of a single set of scale values.

DISCUSSION

Torgerson (1961) concluded that if the subject appreciates only a single relation between a pair of

stimuli, it would not be possible to test empirically between distance and ratio interpretations of this relation. 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.

Scale values defined by the subtractive model for both two-stimulus tasks agree with those derived from the ratio of differences model applied to data obtained from a "ratio of differences" task and with scales derived from a difference of differences model applied to the other three four-stimulus tasks. Since the results of the four-stimulus experiments interlock with the two-stimulus results, it appears that the single comparison process for a pair of stimuli can best be represented by subtraction. The premises of the subtractive theory (P1-P7) are consistent with these results.

The present data fit well in a mosaic of consistent findings which suggest that, for a variety of continua, judges compare two stimuli by subtraction whether instructed to rate "differences" or estimate "ratios." The subtractive theory has the following advantages over the ratio theory: (1) it yields scale values that agree with scales defined by range-frequency theory (Rose & Birnbaum, 1975); (2) it yields scales for easterliness and westerliness that are linearly related (Birnbaum & Mellers, 1978); (3) it predicts that subjects can judge both ratios and differences of stimulus *intervals* (Birnbaum, 1978; Veit, 1978); and (4) it yields consistent scale values for a variety of comparison tasks. The ratio model, without modification, cannot give a consistent account of these findings.

One could ask the question, "can a ratio theory be saved for the present data by replacing the operation of subtraction with division throughout Premises P2-P7?" The answer is that an exponential transformation of all of the models would yield a set of equations that would reproduce the rank orders of the data equally well. However, the "ratio of differences" task would be represented by the following equation:

$$RD_{ijkl} = J_{RD}\left\{\exp[(s_j - s_i)/(s_l - s_k)]\right\}$$

$$= J_{RD}\left\{[\exp(s_j - s_i)]^{\frac{1}{s_l - s_k}}\right\}$$

$$= J_{RD}\left\{[s_j^*/s_i^*]^{\frac{1}{s_l - s_k}}\right\}, \quad (4)$$

where $s^* = \exp(s)$. Equation 4 not only violates the scale convergence criterion within itself, requiring two different scales, s and s^*, but also 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"). This 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 Figure 2 requires that J_{RD} be logarithmic for "ratios of differences." On the basis of simplicity, the subtractive theory is preferable to this modified ratio theory.

The subjective likeableness of adjectives may be inherently no more than an interval scale, like projections on a line in subjective space (Birnbaum, 1978; Veit, 1978). If so, intervals are meaningful but ratios are not. Only when there is a well-defined zero point, as in the case of judging relations between differences, may judges actually compute ratios. This interpretation would explain why subjects could compute "ratios of differences" and "differences of differences," but not "differences of ratios" or "ratios of ratios."

Conclusions

The data obtained from six tasks 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 approximate 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 best represented by subtraction.

REFERENCES

ANDERSON, N. H. Likeableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, 1968, **9**, 272-279.

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 & Psychophysics*, 1974, **15**, 89-96. (b)

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

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

BIRNBAUM, M. H., & MELLERS, B. A. Measurement and the mental map. *Perception & Psychophysics*, 1978, **23**, 403-408.

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

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 anaysis of variance. *Behavioral Science*, 1969, **14**, 165-166.

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

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

PARDUCCI, A., & PERRETT, L. F. Category rating scales: Effects of spacing and frequency of stimulus values. *Journal of Experimental Psychology*, 1971, **89**, 427-452.

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

ROSE, B. J., & BIRNBAUM, M. H. Judgments of differences and ratios of numerals. *Perception & 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.

STEVENS, S. S., & GALANTER, E. H. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, **54**, 377-411.

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.

VEIT, C. T. Ratio and subtractive processes in psychophysical judgment. *Journal of Experimental Psychology: General*, 1978, **107**, 81-107.

NOTES

1. Quotation marks are used throughout to denote instructions to judge "ratios," "differences," "ratios of differences," etc., or numbers obtained with such judgments; quotation marks are not used for the statements about models (e.g., ratio model, ratio of differences model) or theoretical statements about actual ratios and differences.

2. In previous experiments (Rose & Birnbaum, 1975; Veit, 1978) and pilot work, it has been learned that some care is required in teaching these tasks to the subjects. One can always ask, do the judges understand the task? In pilot work, subjects were trained on the algebra of ratios and differences, using the procedure of Rose and Birnbaum (1975, Experiment 2). They were taught that reversing the order of stimuli produces a reciprocal ratio but a minus difference. They were taught that although the differences between 1 and 2, 2 and 3, 3 and 4, etc., are all the same, the ratios change: 1/2, 2/3, 3/4, etc. Although the ratios 1/2, 2/4, 4/8 are the same, the differences vary, $1-2=-1$, $2-4=-2$, $4-8=-4$. The judges were taught subtraction of lengths by cancellation. Graphs illustrated that the difference between two lengths is the length that remains after one length has been subtracted out. Ratios of lengths were also taught graphically, the ratio representing the number of times one length can partition the other. Those judges made numerical calculations to prove that they understood, at least intellectually, the distinction between ratios and differences. The test included numerical ratios greater than and less than one, and included positive, zero, and negative differences. Subjects were able to perform the numerical (intellectual) tasks without error. Rose and Birnbaum (1975) found no discernible difference between the data of subjects who received this training and those who did not, given the subjects showed minimal evidence of following instructions. In spite of this training, some of the subjects in the pilot experiment with adjectives failed to follow instructions adequately for the complex, four-stimulus tasks. The data were similar to those reported here, but it was deemed necessary to repeat the experiments using greater care to insure that the judges understood the instructions.

In the experiments reported here, judges were given additional training on the response procedures of the tasks and were tested on trials that would assess understanding of the instructions, without requiring assumptions about the scale values or model. The test trials given the first day contained cases where a pattern was implied on the basis of the rank order of the scale values. For example, the "ratio of differences" for (honest – cruel)/(shy – cruel) is expected to be greater than 100, since the denominator is expected to be a smaller difference. The warm-ups were designed to contain examples in each response range between special cases for each task. Training trials were checked for special cases and patterns (e.g., "ratios of ratios" should decrease as the second ratio increases). These procedures trained the subjects sufficiently so that all were able to produce responses to the special cases that were superficially consistent with the instructions for the task.

3. Medians have the attractive property that the rank order of medians is invariant under monotonic transformation of the raw scores.

(Received for publication December 7, 1977;
revision accepted May 1, 1978.)