Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement.* Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 401-485.

17 Controversies in Psychological Measurement

Michael H. Birnbaum
University of Illinois
Urbana-Champaign

Psychological measurement is the oldest area of scientific research in psychology and probably the area with the most sophisticated controversies. The chapters in this volume reflect a good deal of disagreement on the fundamental question: How should we measure subjective values? This chapter reviews some of these controversies and presents new points of view on some old, but unsettled, problems.

1. There are two popular methods for obtaining "direct measures" of psychological value—the methods of category rating and magnitude estimation. If category ratings and magnitude estimations are linearly related to subjective value, they should be linearly related to each other. Instead, magnitude estimations are often a positively accelerated function of category ratings. This apparent contradiction has long troubled psychologists, and several theories have been proposed to explain the discrepancy. Section A notes that the relationship between ratings and magnitude estimations varies because both depend on stimulus spacing and the range of responses implied in the instructions. Therefore, theories that assume an invariant relationship between the results of the two procedures face grave difficulties.

2. Because the instructions for magnitude estimation (**M**) seem to focus on "ratios" whereas the instructions for ratings (**C**) seem to focus on "intervals," it seems reasonable to speculate that the relationship between **C** and **M** could be better understood in terms of the comparison processes of the judge. Section B reviews experiments designed to test the hypothesis that judges use the same comparison operation despite instructions to judge "differences" or "ratios." (In this chapter, quotation marks are used to denote the instructions given to the subject or the

responses obtained with such instructions. It is possible to empirically test the hypothesis that "ratio" judgments, for example, fit a ratio model, so it is important to maintain the distinction between the task given the subject and the model used to represent the data.)

3. Many experiments, reviewed in Section B, are consistent with the hypothesis that subjects use the same comparison process to judge both "differences" and "ratios." If subjects use only one operation, is there any way to decide empirically how to represent that operation? Section C discusses more general theories of stimulus comparison that make predictions for tasks in which judges are asked to compare two stimulus relations, for example, to judge the "ratio of two differences" or the "difference between two ratios." In this wider realm, it is possible to test among theories that would otherwise be impossible to discriminate. Evidence from three studies suggests that the "basic" operation for comparing two stimuli is subtraction.

4. Contextual effects in scaling are discussed in Sections A, D, and E. In Section A, contextual effects due to stimulus spacing in category ratings and magnitude estimations are shown to be comparable in form for the two procedures. However, Section D shows that in stimulus-comparison experiments, it may be possible to derive scales that are largely independent of stimulus distribution. In certain situations it is possible to localize the effects of stimulus distribution in the final stage of processing (i.e., in the response function). Section D also presents evidence that in cross-modality comparisons a stimulus is compared in relation to other stimuli within its own modality, and contextually determined values within modality are compared between modalities. Hence, scale values derived from cross-modality comparison depend on the stimulus contexts.

5. Section E discusses philosophical implications of contextual effects for methodology. Some have argued that there is a "right" way to do psychophysical experiments and have advocated experimental designs that would preclude evaluation of the theories upon which the methodology is based. An alternative point of view is presented in which contextual effects are regarded as basic to studies of scaling, and they are therefore accepted and even welcomed.

6. Section F takes up controversies in measurement and model testing. The parallelism test of functional measurement is shown incapable of simultaneously establishing the validity of the response scale and model. Two areas of research, impression formation and the size-weight illusion, are reviewed to challenge previous conclusions of functional measurement and to show how methodological loopholes in simplistic application led to inappropriate conclusions. Improved techniques for model testing are discussed.

7. Section G evaluates related theories of psychophysics that attempt to encompass a wide array of data. It is shown that theories requiring different scales of sensation for different tasks are not yet needed by the data and simpler theories that assume a single scale of sensation remain consistent with a variety of data.

A. JUDGMENT FUNCTIONS IN SINGLE STIMULUS EXPERIMENTS

The overt response, be it a category judgment, magnitude estimation, linemark, physical estimate, cross-modality match, or physical adjustment, depends on the context: the stimulus range, spacing, frequency of presentation, etc. These effects cannot be "avoided" and should not be ignored either at the practical or theoretical level. One way to represent contextual effects is to express the overt response as a function of the subjective value of a stimulus, where the function is permitted to depend on the contextual features of the experiment.

Category Ratings vs. Magnitude Estimations

Let $\mathbf{C}_{ik}$ and $\mathbf{M}_{ik}$ be the category judgment and magnitude estimate of stimulus Φ_i in context k, having subjective value s_i. One can then express the judgments as follows:

$$\mathbf{C}_{ik} = J_{\mathbf{C}_k}(s_i); \tag{A.1}$$

$$\mathbf{M}_{ik} = J_{\mathbf{M}_k}(s_i); \tag{A.2}$$

where $J_{\mathbf{C}_k}$ and $J_{\mathbf{M}_k}$ are the strictly monotonic judgment functions for context k. Equations A.1 and A.2 make clear the distinction between subjective value, s, and the overt response. If the modulus in magnitude estimation or the number of categories in category rating were changed, the overt judgments would change, but one would not want to conclude that the sensations changed. The subscripts (k) for context include any change in procedure for responding that is likely to influence the judgment (or output) function. It seems reasonable to suppose that the functions $J_{\mathbf{C}_k}$ and $J_{\mathbf{M}_k}$ lawfully depend upon such contextual features as the number of categories, stimulus spacing, modulus, etc.

The relationship between category ratings and magnitude estimates in contexts k and n can be expressed as follows:

$$\mathbf{M}_{ik} = J_{\mathbf{M}_k}[J_{\mathbf{C}_n}{}^{-1}(\mathbf{C}_{in})] \tag{A.3}$$

where $J_{\mathbf{M}_k}(J_{\mathbf{C}_n}{}^{-1})$ represents the relationship between ratings and magnitude estitimates of the same stimuli.

At one time it was thought that one could operationally define judgments as in Eqs. A.1 and A.2 to be "direct" measures of sensation, a definition that corresponds to assuming $J_{\mathbf{M}}$ and/or $J_{\mathbf{C}}$ are linear. However, category ratings and magnitude estimations are typically nonlinearly related (Stevens & Galanter,

1957). Therefore, if s is the same in both equations, then $J_{\mathbf{C}}$ and $J_{\mathbf{M}}$ cannot both be linear. Typically, **M** is positively accelerated relative to **C**. Torgerson (1960) noted that log **M** is often linearly related to **C**. Eisler (1962) found that log (**M** + b) may be more nearly linear to **C**. Marks (1974, 1979), Orth (this volume), Wegener (this volume), and others have represented this relationship with linear functions of power functions, $\mathbf{C} = a\mathbf{M}^b + d$, where a, b, and d are arbitrary constants.

Although **M** is typically a positively accelerated function of **C**, the relationship between **M** and **C** changes as a function of contextual details of the experiments, as is illustrated in the next section.

Contextual Effects in "Direct" Scaling

The following experiment illustrates typical findings. Groups of subjects were instructed to judge the darkness of the dot patterns in either half of Fig. 17.1 using either category rating or magnitude estimation. There were eight different groups of subjects. Four groups made category ratings and four groups made magnitude estimations. Within either response procedure, two groups of subjects received stimuli spaced according to a positively skewed stimulus distribution (relative to logΦ), as shown in the left side of Fig. 17.1, and two different groups of subjects received stimuli spaced according to a negatively skewed distribution, as on the right of Fig. 17.1. Note that both distributions have six values in common. Patterns labeled 9, 11, 6, 10, 1, and 7 have 12, 18, 27, 40, 60, and 90 dots, respectively, in both contexts. If there were no effects of the other stimuli presented for judgment, then these common stimuli should receive the same judgments in both contexts.

Category Ratings. For the category-rating experiments, two groups used a five-point scale in which 1 = *lightest* pattern and 5 = *darkest* pattern. The other two groups were given a 1–100 scale with the end points anchored to the end stimuli in the same way.

Results for the category-rating tasks are shown in Fig. 17.2. The upper panel shows the results for the 1–100 scale, and the lower panel shows the results for the 1–5 scale. The two curves within each panel show that mean ratings can be either positively accelerated or negatively accelerated relative to logΦ, depending on the spacing of the stimuli chosen for judgment. The general shape of the trends is consistent with Parducci's range-frequency theory. Parducci (this volume) has shown that the magnitude of the contextual effect due to stimulus distribution decreases with increasing number of response categories and increases as a function of the number of stimulus levels. The present data, obtained with 11 stimulus values, show that the contextual effect for the 100-point rating scale remains quite large.

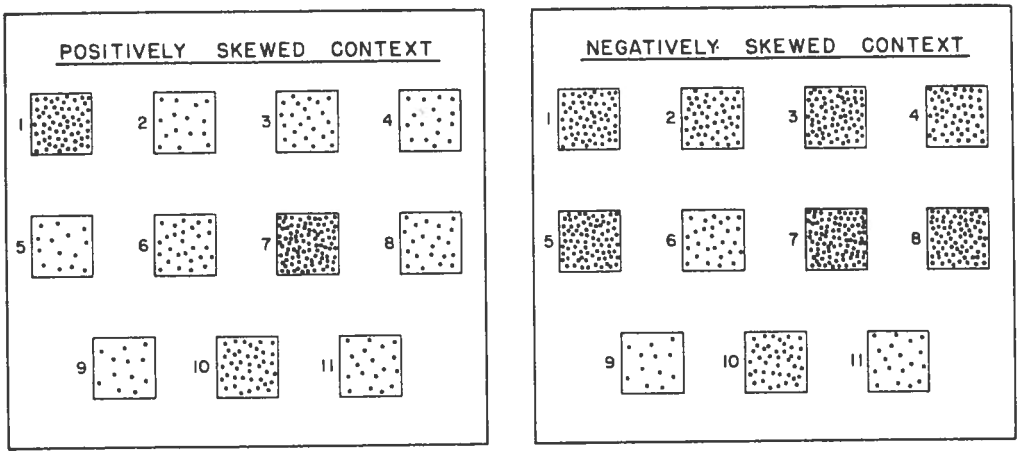


FIG. 17.1. Two stimulus distributions (contexts). Subjects judged darkness of each dot pattern. Different groups of subjects received different stimulus distributions. Note that patterns numbered 9, 11, 6, 10, 1, and 7 are identical in both contexts; these stimuli have 12, 18, 27, 40, 60, and 90 dots, respectively. From Birnbaum and Mellers (1980a).

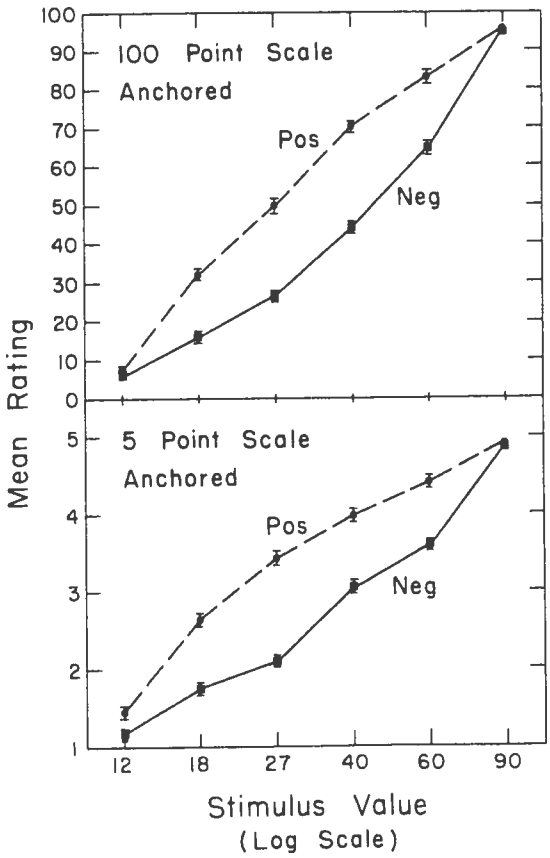


FIG. 17.2. Contextual effects in category ratings. Mean ratings are plotted against stimulus values, spaced on abscissa in log scale. Dashed curves show ratings when stimuli are spaced according to the positively skewed stimulus distribution in Fig. 17.1; solid curves are for the negatively skewed distribution. Upper panel shows results for 1-100 rating scale, lower panel for 1-5 rating scale. Brackets show plus and minus one standard error. Trends are in direction predicted from Parducci's range-frequency theory. From Birnbaum and Mellers (1980a).

Magnitude Estimations. For the magnitude-estimation experiments, all subjects were told to call the lightest pattern "100" and to assign numbers to each stimulus so that the ratios of the numbers would match the "ratios of the subjective darkness of the sensations." Both groups were encouraged to use whatever numbers they wished, but different examples were given in the instructions to help explain the task. In one case, the examples went as high as "300" (if the pattern seems *three times as dark* as the lightest pattern, say 300). In the other case, the examples went as high as "900" (if the pattern seems *nine times as dark*). Note that this change in the magnitude-estimation instructions is subtle; according to early theories of magnitude estimation, this aspect of the instructions should theoretically have no effect. Instead, Fig. 17.3 shows that it has a great effect.

The magnitude estimations of the common stimuli are shown in Fig. 17.3, with a separate curve for each condition. If the stimulus values chosen had no effect on magnitude estimations, and if the examples used in the instructions had no effect, then all four curves should coincide. Instead, the difference between the open and solid points shows that when the examples range as high as "900," the subjects use numbers that average much higher than when the largest example is only "300." Inasmuch as the exponent obtained in a magnitude-estimation experiment depends largely on the (log) response range, it appears that the exponents obtained in magnitude-estimation studies may relate more closely to the experimenter's range of examples than to the subjects' range of sensations. Robinson (1976) and Poulton (1979) reached similar conclusions.

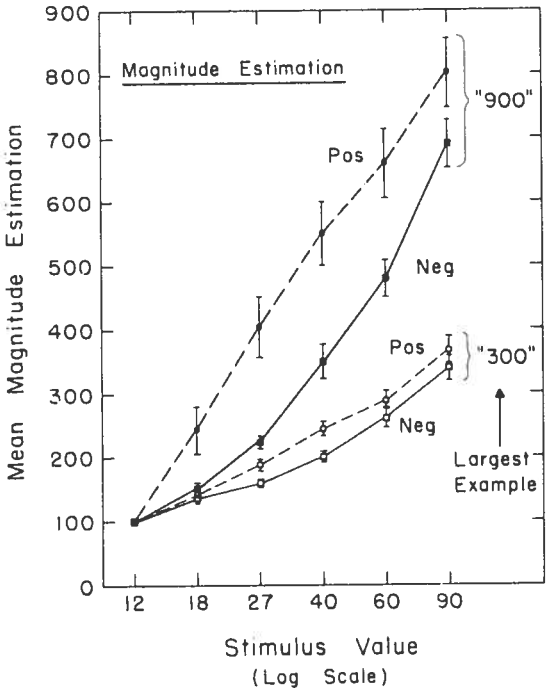


FIG. 17.3. Contextual effects in magnitude estimation. Mean magnitude estimations are plotted against stimulus values as in Fig. 17.2. The 12-dot stimulus was the standard and was assigned the value "100." Dashed lines show results when stimuli were spaced in a positively skewed distribution; solid lines are for negatively skewed conditions. Upper two curves show results when instructions included an example response as high as "900"; lower two curves show results when largest example was "300." Brackets show plus and minus one standard error. If there are no contextual effects in magnitude estimation, the curves should all coincide. From Birnbaum and Mellers (1980a).

Furthermore, Fig. 17.3 shows that magnitude estimations depend on the stimulus spacing. Magnitude estimations appear to show similar contextual effects to those for category ratings. Parducci (1963) found similar results.

Quite clearly, because the relationships between **M** and Φ and between **C** and Φ can have so many different forms, the relationship between **M** and **C** should not be theorized to be an invariant functional form.

Contextual effects due to stimulus distribution cannot be adequately approximated by power functions. By increasing the stimulus density in the center of the array or at the ends, the functions can be changed in their cubic components (sigmoidal trends). Therefore, the relationship between category ratings and magnitude estimations cannot be adequately represented by power functions, even allowing different exponents for different contexts. Because the relationship between stimulus and response can attain so many different functional forms, Nihm (1976) has suggested (satirically) that the power function be replaced by the polynomial, which given enough parameters can describe any finite set of data.

Reactions to Contextual Effects

Some investigators, having realized the existence of contextual effects, have reacted by adopting extreme positions. Some of these are discussed by Poulton (1968, 1979). One reaction is that contextual effects are undesirable and should be "avoided," averaged away, or ignored. In order to "avoid" contextual effects, it has been argued that everyone should use the same stimulus distribution, response examples, and so on. Then these different functions (as in Fig. 17.3) would not trouble us. But then, how do we decide which context (stimulus distribution, response procedure, etc.) is the "right" one? Will we not be accused of choosing the context to produce the desired effect?

The second reaction, which is also unfortunate, is that if the functional relationship between stimulus and response can be manipulated, then one cannot assume any metric properties in the response. It could be thought that because the response with one procedure (rating) is nonlinearly related to the response with another (estimation), one cannot assume any more than that the response is an unknown and perhaps unpredictable monotone function of subjective value. It is argued in sections D and E that this reaction is too pessimistic and ignores the lawfulness of the effects in Figs. 17.2 and 17.3. Instead, it can be argued that any complete theory of psychophysics must give an account of the response procedure and contextual effects in order to explain the lawful numerical changes that result as a function of these variables.

The effects of stimulus range, stimulus spacing, and frequency are best understood for category ratings and are well-described by Parducci's range-frequency theory (Parducci, 1963, 1965, 1974, this volume; Parducci & Perrett, 1971).

Indeed, the lawfulness of the stimulus-spacing effect can be used to define a psychophysical scale (Birnbaum, 1974c).

Range-Frequency Theory

A general form of Parducci's range-frequency theory can be written as follows:

$$\mathbf{C}_{ik} = a_k \left[\frac{s_i}{s_m - s_o} \right] + b_k G_k(s_i) + c_k \tag{A.3}$$

where $\mathbf{C}_{ik}$ is the category rating of stimulus i in context k; s_i is the subjective value of the stimulus; s_m and s_o are the subjective values of the maximum and minimum stimuli in context k; and $G_k(s_i)$ is the (cumulative) proportion of subjective values less than s_i in context k. The linear constants, a_k, b_k, and c_k reflect the weight of the range and frequency principles and may depend on the number of stimulus levels and the number of categories (Parducci, this volume).

Birnbaum (1974c) noted that when the stimulus and response ranges are held constant, and the stimuli are presented simultaneously, ratings can be well-approximated by the model:

$$\mathbf{C}_{ik} = as_i + bF_k(\Phi_i) + c \tag{A.4}$$

where $F_k(\Phi_i)$ is the cumulative proportion of stimuli less than Φ_i in context k; and a, b, and c are constants. It follows that

$$as_i + c = \mathbf{C}_{ik} - bF_k(\Phi_i) \tag{A.5}$$

Thus, because F_k is known, range-frequency theory provides a basis for estimating scale values. Instead of "avoiding" contextual effects by holding the stimulus distribution fixed to some arbitrary value, Birnbaum (1974e) argued that the systematic manipulation of the context allows one to test theories, such as range-frequency theory, and simultaneously estimate context-free scale values. This issue is taken up in greater detail in Section E.

B. "RATIOS" AND "DIFFERENCES"

Torgerson (1961) postulated that the contradiction between magnitude estimations and category ratings might be explained by the premise that judges perceive only a single relation between a pair of stimuli, irrespective of instructions to judge "differences" or "ratios." This conjecture, which could not be tested in the early research, has received new support from recent studies that have independently manipulated stimulus levels (for reviews, see Birnbaum, 1978, 1979, 1980a). The following empirical findings have emerged from this research:

1. With certain experimental methods, magnitude estimations of "ratios" closely fit the ratio model. The raw data, when plotted against the estimated scale value of the comparison stimulus with a separate curve for each standard, show the appropriate pattern of bilinearity predicted by the ratio model.
2. Category ratings of "differences" fit the subtractive model, showing approximate parallelism, when the data are plotted in the same way.
3. Scale values derived from the fit of the ratio model applied to "ratios" are very close to an exponential function of scale value derived from the fit of the subtractive model applied to "difference" judgments.
4. Judgments of "ratios" and "differences" are monotonically related. These empirical findings are consistent with the hypothesis that the same operation and scale values underlie both procedures.

Theories of Ratios and Differences

Two-Operation Theory. According to this theory, subjects perform both tasks using two operations on the same scale values. "Ratio" judgments are given by the equation:

$$\mathbf{R}_{ij} = J_{\mathbf{R}}[s_j/s_i] \tag{B.1}$$

where $\mathbf{R}_{ij}$ is the "ratio" judgment of stimulus j relative to i, and $J_{\mathbf{R}}$ is the monotonic judgment function. "Difference" judgments are given by the equation:

$$\mathbf{D}_{ij} = J_{\mathbf{D}}[s_j - s_i] \tag{B.2}$$

where $\mathbf{D}_{ij}$ is the "difference" response, and $J_{\mathbf{D}}$ is the judgment function for "differences."

This theory implies that $\mathbf{R}_{ij}$ and $\mathbf{D}_{ij}$ should *not* be monotonically related, in general, but instead that the rank orders of these matrices should be different but appropriately interrelated (Krantz, Luce, Suppes, & Tversky, 1971). For example, as a constant difference is moved up the scale (e.g., $2 - 1 = 3 - 2 = 4 - 3 = 5 - 4$, etc.), the corresponding ratios approach 1 ($\frac{2}{1} > \frac{3}{2} > \frac{4}{3} > \frac{5}{4}$, etc.). As a constant ratio is moved up the scale (e.g., $\frac{2}{1} = \frac{4}{2}$), absolute differences increase ($2 - 1 < 4 - 2$).

The left side of Fig. 17.4 plots actual ratios against actual differences for a 7×7, A by B, factorial design, using successive integers from 1 to 7 as levels of A and B. The ordinate plots A/B, the abscissa plots A − B, and separate curves connect points with the same value of B (curve parameters). The highest curve (solid points) plots A/1 vs. A-1. The curve with the lowest slope (solid diamonds) plots A/7 vs. A-7. Note that the relationship between actual ratios and differences cannot be expressed by any function of a single variable because for

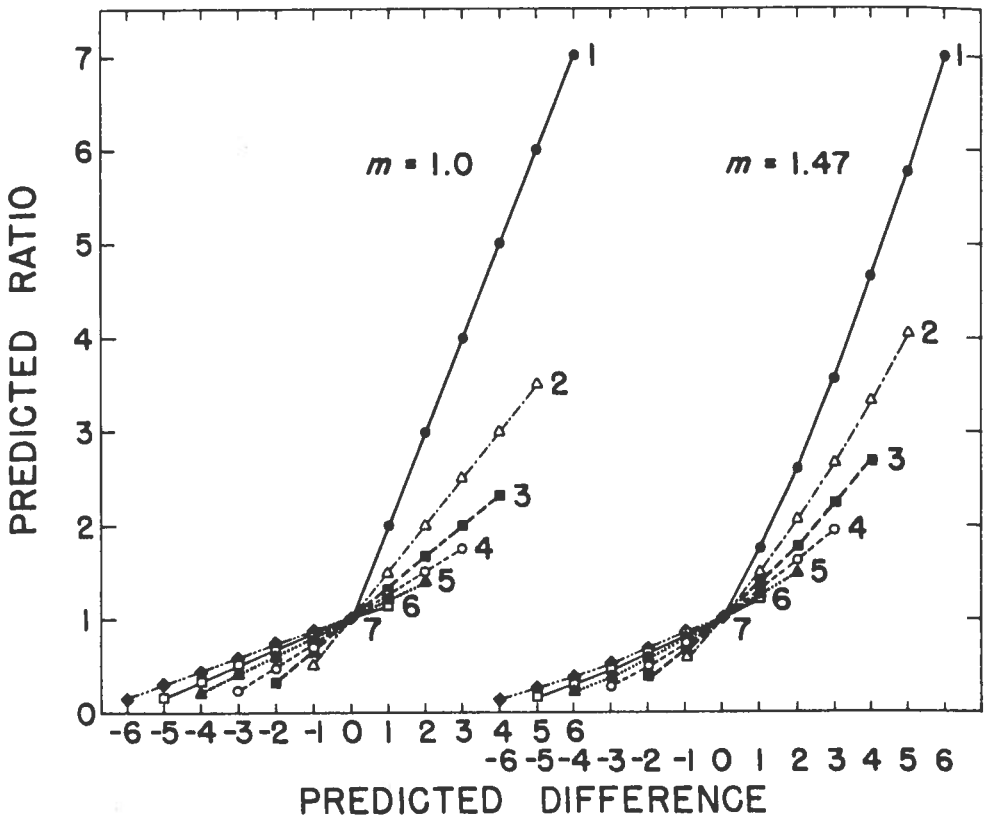


FIG. 17.4. Theoretical predictions of theory that judgments of "ratios" and "differences" are governed by ratio and difference operations. Left: A/B plotted against A−B with a separate symbol (and separate curve) for each value of B. Predictions for 7 × 7 design using successive integers from 1 to 7 for A and B. Right: $(A/B)^{1.47}$ plotted against 2.17(A−B) in the same fashion, except the seven values of A and B were evenly spaced between 1 and 3.76 (i.e., 1 to $7^{.68}$). From Birnbaum (1980a).

any given difference there does not exist one unique ratio, but many, and for any ratio there exist many differences.

The judgment functions, J_R and J_D, serve to monotonically stretch the ordinate and abscissa of Fig. 17.4, but it should be clear that as long as A and B have been independently manipulated over a sufficient range, the ordinal pattern of ratios vs. differences of Fig. 17.4 should remain.

The right side of Fig. 17.4 shows the expected relationship between ratios and differences if the subjective stimulus range is only $7^{.68}$, or 3.76, and if the J_R function is a power function with an exponent of 1.47. The abscissa shows differences (times 2.17), and the ordinate shows ratios raised to the 1.47 power. Thus, the ordinate shows $(A/B)^{1.47}$, plotted against 2.17 (A − B) on the abscissa, using seven levels of A and B spaced evenly between 1.0 and 3.76.

This smaller stimulus range was chosen so that the largest "ratio" (7) would be consistent with typical results from "ratio" experiments, given the average output exponent (1.47) reported by Rule and Curtis (this volume). Thus, if

magnitude estimations of "ratios" are a power function of subjective ratios with an exponent in the range of values reported by Rule and Curtis, "ratios" and "differences" should be quite distinct and have different rank orders as shown on the right of Fig. 17.4.

One-Operation Theory. If subjects use only one operation for both tasks, and if that operation can be represented by subtraction, then the data can be represented by the following:

$$\mathbf{R}_{ij} = J_{\mathbf{R}}[s_j - s_i] \tag{B.3}$$

$$\mathbf{D}_{ij} = J_{\mathbf{D}}[s_j - s_i] \tag{B.4}$$

where $J_{\mathbf{R}}$ represents the strictly monotonic judgment function for magnitude estimations of "ratios" and $J_{\mathbf{D}}$ represents the strictly monotonic judgment function for ratings of "differences." It follows that $s_j - s_i = J_{\mathbf{D}}^{-1}[\mathbf{D}_{ij}]$. Therefore, $\mathbf{R}_{ij} = J_{\mathbf{R}}[J_{\mathbf{D}}^{-1}(\mathbf{D}_{ij})]$. Because $J_{\mathbf{R}}J_{\mathbf{D}}^{-1}$ is monotonic, one-operation theory implies that "ratios" are monotonically related to "differences."

With the experimental procedures used in the research reviewed by Birnbaum (1980a), it has been found that the $J_{\mathbf{R}}$ function for magnitude estimation can be well-approximated by an exponential function, and $J_{\mathbf{D}}$ for ratings can be approximated by a linear function. In this case, the model can be written:

$$\mathbf{R}_{ij} = a_{\mathbf{R}}\exp[c_{\mathbf{R}}(s_j - s_i)] + b_{\mathbf{R}}, \tag{B.5}$$

$$\mathbf{D}_{ij} = a_{\mathbf{D}}(s_j - s_i) + b_{\mathbf{D}}, \tag{B.6}$$

where $a_{\mathbf{R}}$, $a_{\mathbf{D}}$, $b_{\mathbf{R}}$, $b_{\mathbf{D}}$, and $c_{\mathbf{R}}$ are constants. The comparison operation is subtraction in both cases. It follows that $\mathbf{R}_{ij}$ should be exponentially related $\mathbf{D}_{ij}$.

A Brief Review of Research on "Ratios" and "Differences"

Nine experiments that obtained "ratio" and "difference" judgments are summarized in Fig. 17.5. "Ratios" are plotted on the ordinate against "differences" on the abscissa, with separate symbols for each divisor, as in Fig. 17.4. Instead of resembling the predictions in Fig. 17.4 of the two-operation theory, the data appear more closely to fall on a single monotone function in each case. "Ratios" are roughly an exponential function of "differences," as shown by the resemblance of the data to the exponential curves, which have been fit through just two points (0, 1) and the highest point for each set of data.

Figure 17.5 shows that for experiments with heaviness (Birnbaum & Veit, 1974a), pitch of pure tones (Elmasian & Birnbaum, 1979), darkness of dot patterns (Birnbaum, 1978), darkness of grays (Veit, 1978), loudness of 1000 Hz tones (Birnbaum & Elmasian, 1977), likeableness of adjectives (Hagerty & Birnbaum, 1978), and easterliness or westerliness of U.S. cities (Birnbaum &

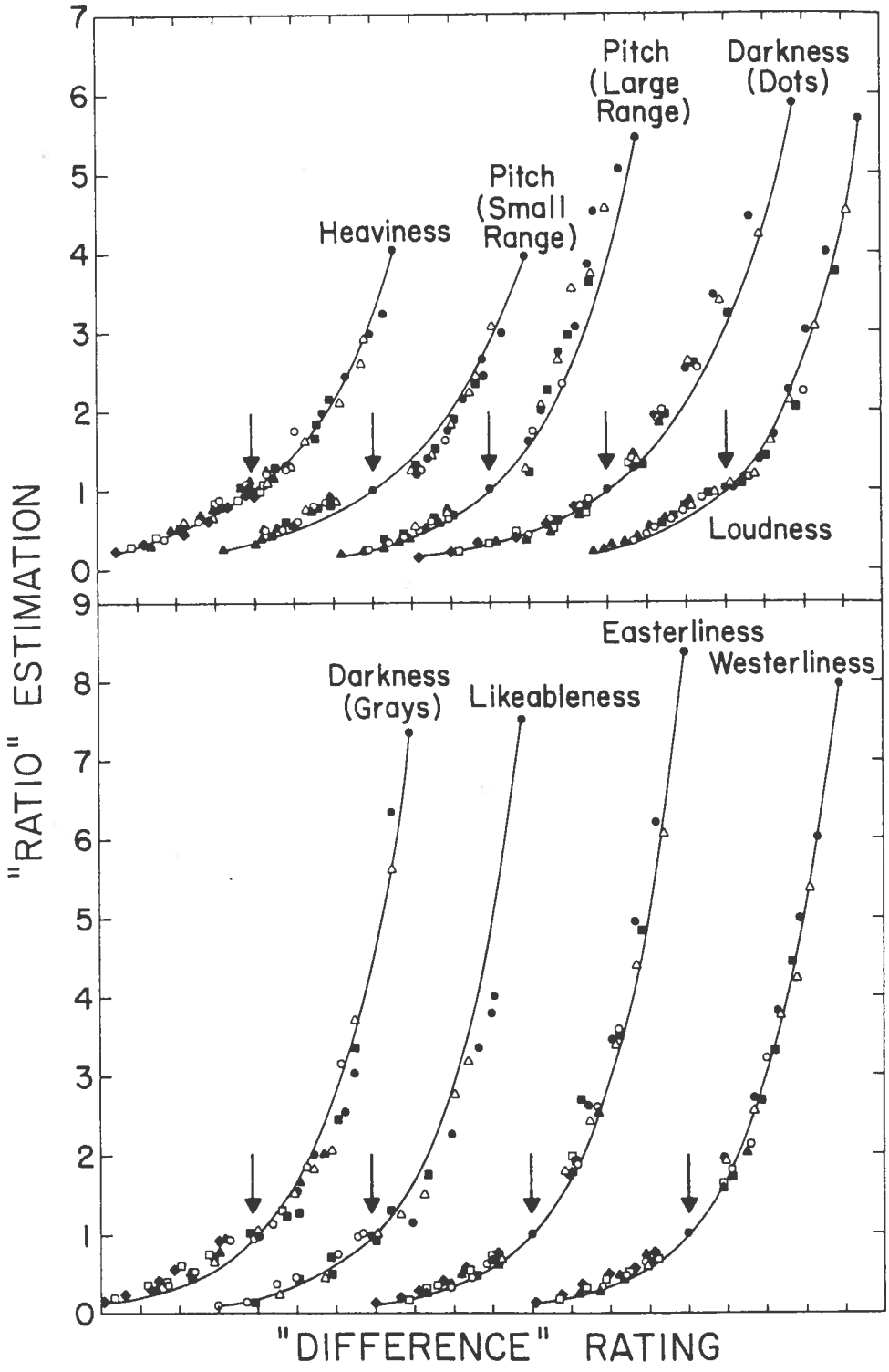


FIG. 17.5. Judgments of "ratios" and "differences," plotted as in Fig. 17.4, for nine experiments. Magnitude estimations of "ratios" are nearly a monotonic function of ratings (or estimations) of "differences," unlike the predictions of Fig. 17.4, but consistent with the theory that the same operation underlies both tasks. Exponential functions have been drawn through the point corresponding to a difference of "zero" and a ratio of "one" (arrows) and the highest point for each experiment. From Birnbaum (1980a).

Mellers, 1978), it appears that "ratios" and "differences" are monotonically related. Similar conclusions were reached by Rose and Birnbaum (1975) for numerical comparisons. Although Parker, Schneider, and Kanow (1975) concluded that judges use two operations for length, Schneider, Parker, Kanow, and Farrell (1976) reached the conclusion that "ratios" of loudness are governed by subtraction. Schneider (this volume) summarizes additional research consistent with the hypothesis that the same comparison process is used for judgments of "ratios" and "differences."

In summary, for a number of social and psychophysical continua, judgments of "ratios" and "differences" can be represented by the same comparison operation. If it is assumed that this operation is subtraction, the $J_{\mathbf{R}}$ function (for magnitude estimations of "ratios") can be approximated by the exponential, and the $J_{\mathbf{D}}$ function (for ratings of "differences") is approximately linear.

However, the subtractive representation,

$$\mathbf{R}_{ij} = a_{\mathbf{R}}\exp(s_j - s_i) \tag{B.7}$$

$$\mathbf{D}_{ij} = a_{\mathbf{D}}(s_j - s_i) \tag{B.8}$$

can be replaced by an equivalent ratio representation as follows:

$$\mathbf{R}_{ij} = a_{\mathbf{R}}(s_j^*/s_i^*) \tag{B.9}$$

$$\mathbf{D}_{ij} = a_{\mathbf{D}}\ln(s_j^*/s_i^*) \tag{B.10}$$

where $s^* = \exp(s)$. In other words, judgments of "ratios" and "differences" are consistent with the proposition that the *same* operation underlies both tasks, but they do not permit specification of what that operation might be.

Is it meaningful to ask whether judges are "really" comparing two stimuli by computing a difference or a ratio? The next section discusses a theoretical and methodological framework in which this question can be answered.

C. RESOLUTION OF THE RATIO-DIFFERENCE CONTROVERSY

The finding that judgments of "ratios" and "differences" are monotonically related is consistent with Torgerson's (1961) hypothesis that judges compare two stimuli by the same operation for both tasks. Torgerson (1961) concluded that if only one operation were used, it would not be possible to *discover* whether the operation is a difference or ratio. Whichever representation was chosen would be a "decision, not a discovery."

However, Birnbaum (1978) and Veit (1978) have shown that with a wider array of data involving both stimulus comparisons (A vs. B) and also comparisons of stimulus relations (AB vs. CD), it becomes possible to discriminate among different theories. Consider the stimuli shown in Fig. 17.6. The observer can be asked to judge the "ratio" (**R**) of A to B or the "difference" (**D**) between A

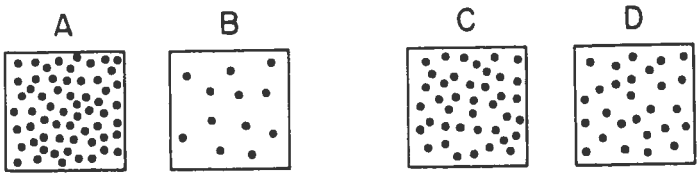


FIG. 17.6. Stimulus array for one trial of a four-stimulus task. From Birnbaum (1980d).

and B, as before. It is also possible to ask the observer to judge the "ratio of the difference" between A and B relative to the difference between C and D, for example. Four of these four-stimulus tasks have been investigated: "ratios of ratios" [**RR** = (A/B)/(C/D)], "ratios of differences" [**RD** = (A − B)/(C − D)], "differences of ratios" [A/B − C/D], and "differences of differences" [(A − B) − (C − D)].

These four-stimulus polynomials can be diagnosed by ordinal and metric analyses (Birnbaum, 1978). Furthermore, by comparing the scale values for the same stimuli across the tasks, the information gained from such an experiment is increased. Veit (1978) employed a "ratio of differences" task in addition to the "ratio" and "difference" tasks for judgments of the darkness of papers of varied reflectance. Hagerty and Birnbaum (1978) used six tasks (**R, D, RR, RD, DR, DD**). (These experiments have been reviewed by Birnbaum, 1978). The next sections illustrate the major findings using data from a new experiment that employed stimuli as in Fig. 17.6, and which replicated the findings of Veit (1978) and those of Hagerty and Birnbaum (1978). The following conclusions have been drawn from this research:

1. Judgments of "ratios of ratios," "differences of ratios," or "differences of differences," can all be represented by the difference of differences model, using the same scale values for the stimuli for all three tasks.
2. However, judgments of "ratios of differences" can be represented by a ratio of differences model.
3. The scale derived from the ratio of differences model is consistent with the scale derived from the subtractive model applied to "difference" and "ratio" judgments.
4. The scale values derived from the fit of the difference of differences model applied to "ratios of ratios," "differences of ratios," and "differences of differences" agree with the scale derived from the ratio of differences model applied to "ratios of differences."
5. The judgment functions for magnitude estimations of "ratios" and "ratios of ratios" can be well-approximated by exponential functions, whereas the other judgment functions are approximately linear.
6. Therefore, the data are consistent with the hypothesis that the basic operation for comparing two stimuli in these continua is subtraction.

Theories of Stimulus Comparison

Figure 17.7 gives an outline for discussing theories of stimulus comparison and combination. In the outline, the physical and subjective values of the stimuli are denoted Φ and s, where $s = H(\Phi)$ is the psychophysical function; the subjective value of a comparison between two stimuli (or the combination of two stimuli) is denoted $\Psi_{ij} = C(s_i, s_j)$, where C represents the comparison (or combination) process. Two comparisons (or combinations) are compared (or combined) by the function $\delta = G(\Psi_{ij}, \Psi_{kl})$; the overt response, **R**, is assumed to be a monotonic function of Ψ in the two-stimulus case and of δ in the four-stimulus case. For comparison with Table 17.1 and Fig. 17.7, let s_j = A, s_i = B, s_l = C, and s_k = D.

Table 17.1 shows five theories of stimulus comparison considered by Birnbaum (1978, 1979). It is useful to consider first the predictions of the theory that judges obey the instructions and use a single scale of subjective value. This theory is labelled Model = Task in Table 17.1.

Figure 17.8 shows calculated ratios and differences for a 7 × 7, A by B, factorial design, using integers from 1 to 7 as in the left of Fig. 17.4. In Fig. 17.8, A/B is plotted on the left as a function of A with a separate curve for each

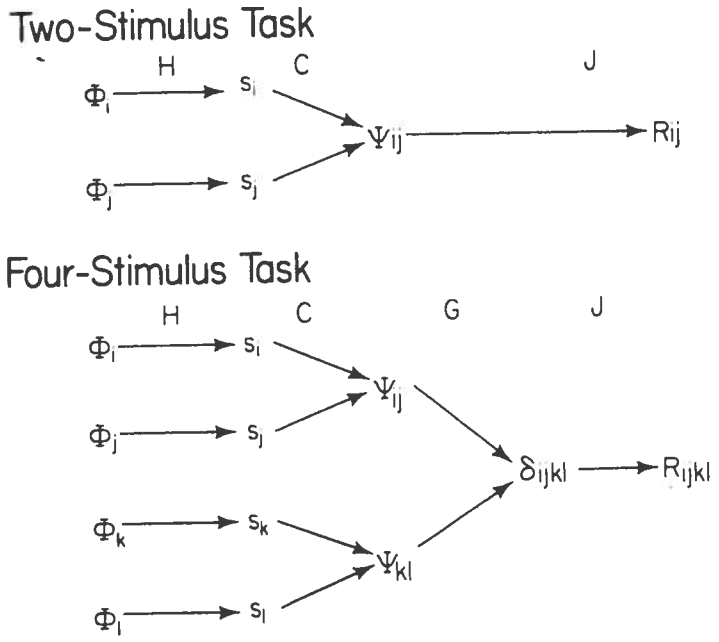


FIG. 17.7. Outline of stimulus comparison and combination for two- and four-stimulus tasks. In the outline, physical values, Φ, are mapped into subjective values, s, by the psychophysical function, H; subjective values are compared (or combined) by the function, $\Psi_{ij} = C(s_i, s_j)$. Two comparisons (or combinations) are compared (or combined) by the function, $\delta = G(\Psi_{ij}, \Psi_{kl})$. The overt response, **R** is assumed to be a monotonic function of the subjective impression, where J represents the judgment function. From Birnbaum (1979).

TABLE 17.1
Theories Discussed by Birnbaum (1978)[a]

	Theory				
Task	*Model = Task*	*Subtractive*	*Ratio*	*Indeterminacy*	*Two-Worlds*
"Ratios"	A/B	A−B	A/B	A−B	a/b
"Differences"	A−B	A−B	A/B	A−B	A−B
"Ratios of Ratios"	(A/B)/(C/D)	(A−B)−(C−D)	(A/B)/(C/D)	(A−B)−(C−D)	$(a/b)/(c/d)$
"Differences of Ratios"	(A/B)−(C/D)	(A−B)−(C−D)	(A/B)−(C/D)	(A−B)−(C−D)	$(a/b)-(c/d)$
"Ratios of Differences"	(A−B)/(C−D)	(A−B)/(C−D)	(A/B)/(C/D)	(A−B)−(C−D)	(A−B)/(C−D)
"Differences of Differences"	(A−B)−(C−D)	(A−B)−(C−D)	(A/B)/(C/D)	(A−B)−(C−D)	(A−B)−(C−D)

[a] A, B, C, D refer to s_j, s_i, s_k, s_l in Fig. 17.7, respectively. Each entry represents the model for each task predicted by each theory. Judgment functions are omitted for simplicity. For the two-worlds theory, $a = \exp(A)$, $b = \exp(B)$, etc.

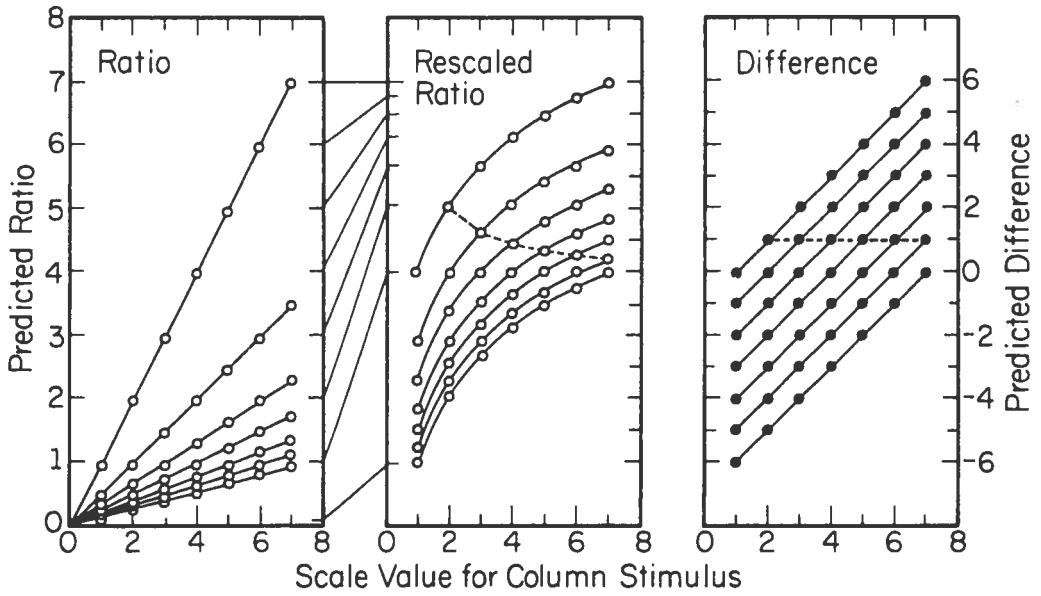


FIG. 17.8. Predicted ratios (A/B) and differences (A-B) for 7-by-7 design using successive integers from 1 to 7 for values of A and B. This figure replots the data of Fig. 17.4 to show bilinearity and parallelism predictions of ratio and subtractive models, respectively. Center panel shows that when ratios are rescaled to parallelism (by the log transformation), the curves would *not* coincide with differences. From Birnbaum (1978).

level of B. On the right of Fig. 17.8, A − B is plotted against A with a separate curve for each level of B. The curves are linearly related to A in each case. They form a bilinear divergent fan for A/B, and the curves are parallel for A − B. The center panel of Fig. 17.8 shows that when ratios are transformed by the log function, log(A/B) = logA − logB, the transformed ratios are parallel, but they are not linearly related to A. Figure 17.8 shows again that A/B and A − B are not monotonically related.

Figures 17.9 and 17.10 show predictions of the theory that Model = Task for the four-stimulus tasks (**RR**, **RD**, **DD**, and **DR**). To compute predictions, the 7 × 7 design (using successive integers from 1 to 7) was factorially combined with a 2 × 2, C by D design, in which the levels of C were five and seven and the levels of D were one and four. The design is thus a 7 by 7 by 2 by 2, A by B by C by D factorial. Therefore, C/D is always greater than 1, and C − D is always greater than 0.

Figure 17.9 shows that the **RR** and **DR** models [(A/B)/(C/D) and A/B − C/D] imply a bilinear interaction between A and B. The **DD** and **RD** models [(A − B) − (C − D) and (A − B)/(C − D)] imply no interaction (parallelism) between A and B. Figure 17.10 shows the form of the A by C by D interactions for the four tasks. Other aspects of these models are discussed by Birnbaum (1978).

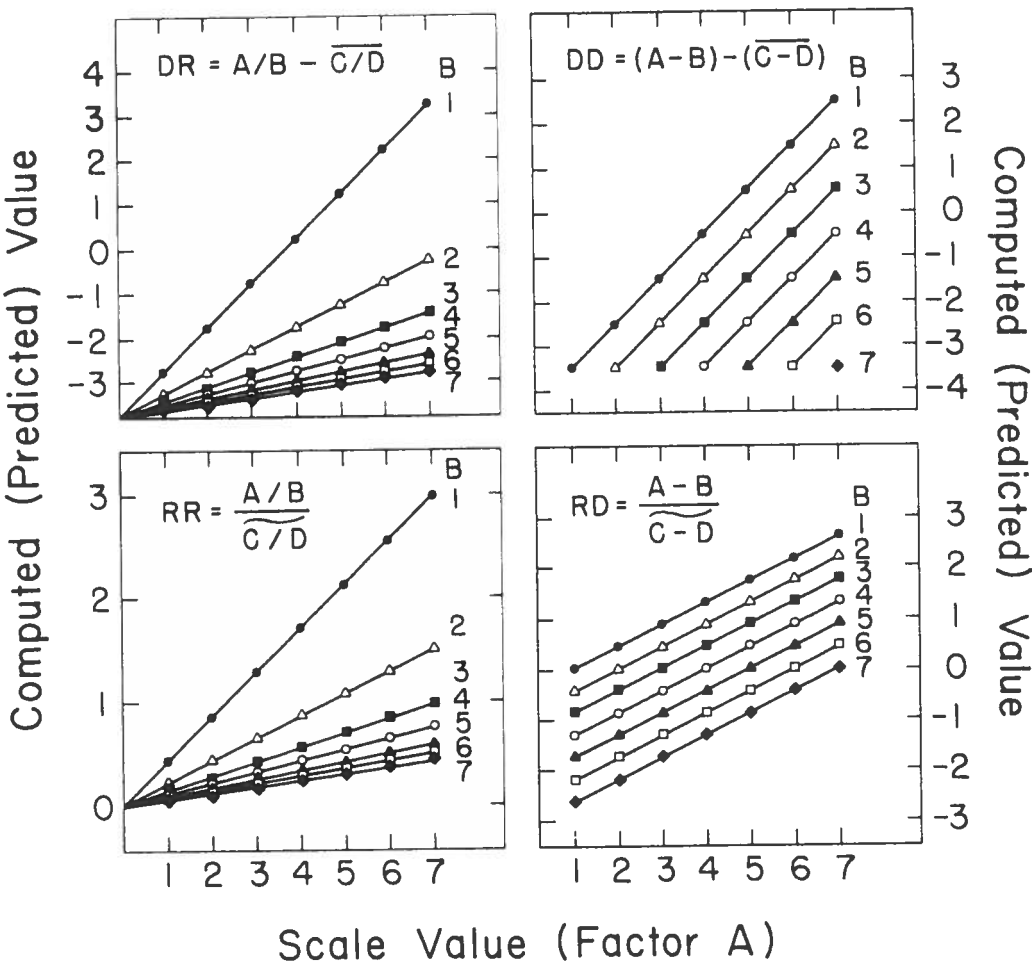


FIG. 17.9. Predicted values for A-by-B cell means, averaged over levels of C and D, for the theory that the model is the same as the task. Note that curves are parallel for ratios of differences (**RD**) and differences of differences (**DD**), and that they show bilinear divergence for ratios of ratios (**RR**) and differences of ratios (**DR**). Note also that curves are all linearly related to scale values of A. From Birnbaum (1980d).

Comparing the Theories

As Birnbaum (1978) and Veit (1978) noted, the four-stimulus polynomials can be distinguished on the basis of ordinal tests analogous to those described by Krantz and Tversky (1971). By adding the extra leverage of the scale convergence criterion, the number of distinct outcomes (and therefore the total constraint of the experiment) is greatly increased.

In addition to the theory that Model = Task, there are four simple theories to consider. The subtractive theory assumes that simple "ratios" and "differences" are computed by subtraction. Once a subjective interval has been computed, however, the subject can compare this interval by either a ratio or difference operation. The ratio theory (comparably) assumes that ratios underlie

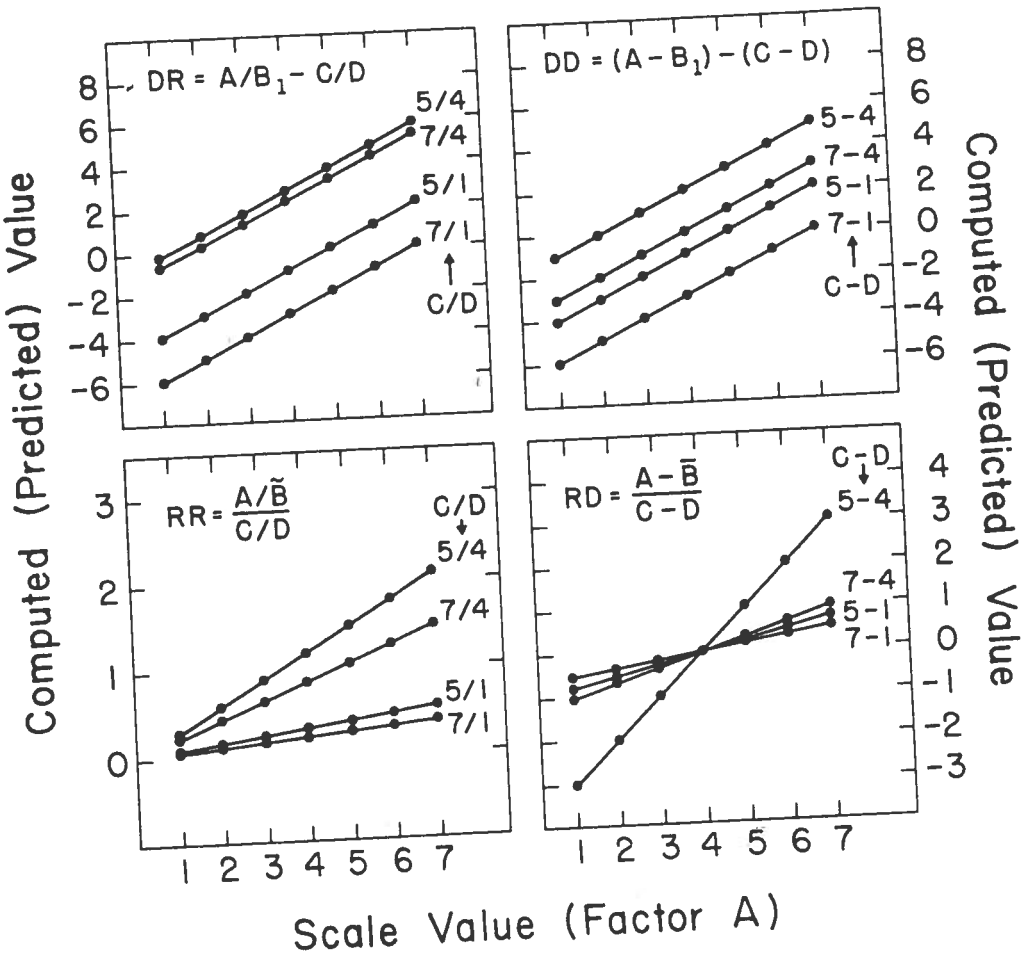


FIG. 17.10. Predicted values for A by C by D, with B fixed to level 1 (for upper panels) and averaged over levels of B (for lower panels) based on the theory that model = task. Note that curves are linearly related to scale values of A in all panels. Note also that A is additive (parallel) with C/D and C-D for **DR** and **DD** models, respectively, and shows a bilinear interaction with C/D and C-D for **RR** and **RD** models, respectively. From Birnbaum (1980d).

the comparison of two stimuli and that judges can compare two ratios by either a ratio or by a difference. The indeterminacy theory assumes that only one operation is possible for comparison of two stimuli or two comparisons. If the data are consistent with this theory, they would not offer any basis for preferring a ratio or subtractive representation. The two-worlds theory assumes that there are two scales and two sets of corresponding operations. The two-worlds outcome could occur if observers in the four-stimulus tasks used arithmetic on their (implicit) judgments of **R** and **D**.

Except for Model = Task, all four theories predict that the **R**, **D**, **RR**, and **DD** data should be rescalable to a difference, or difference of differences, model using the same scale values for A. The theories differ in that the subtractive and two-worlds theories predict that "ratios of differences" should fit the ratio of differences model.

Veit (1978) found that "ratios of differences" of darkness of gray chips could indeed be fit to a ratio of differences model, but could *not* be monotonically rescaled to fit a difference of differences model, as they showed the appropriate violations of joint independence that refute the **DD** or **RR** models but which are characteristic of the **RD** model. Thus, her results were consistent with the subtractive theory but did not test the two-worlds theory.

The ratio theory and two-worlds theory predict that "differences of ratios" should fit the difference of ratios model. Hagerty and Birnbaum (1978) studied all six tasks for the likeableness of adjectives and found no evidence for ratio theory or two-worlds theory. With the assistance of Steven E. Stegner and Bernadette Setiadi, the author has replicated the major findings of Hagerty and Birnbaum (1978) in a new experiment using psychophysical stimuli as in Fig. 17.6.

Experimental Test

In this experiment, 227 undergraduates judged "ratios" and "differences" of the darkness of dot patterns presented in the format of Fig. 17.6 (except without C and D). The number of dots varied from 8 to 90 in seven equal log steps as in the experiment described by Birnbaum (1978, Fig. 4), which used a different stimulus presentation format. After performing both the simple **R** and **D** tasks, each subject served in one of the four-stimulus tasks: **RR**, **RD**, **DR**, or **DD** (there were 41 to 55 different subjects in each condition). The design for these tasks was a $7 \times 7 \times 2 \times 2$, A × B × C × D, in which levels of C were levels five (40 dots) and seven (90 dots) and for D they were one (8 dots) and four (27 dots).

Experimental Results

Figures 17.11, 17.12, and 17.13 show the results for the six tasks, plotted for comparability with Figs. 17.8, 17.9, and 17.10. However, note that data for "ratios," and "ratios of ratios" are plotted against the antilog$_2$ of the estimated scale value, *unlike* Figs. 17.8 through 17.10.

The data were fit to the theories in Table 17.1, with the result that the subtractive theory gave the best overall fit. The predicted curves in Figs. 17.11, 17.12, and 17.13 are based on the following model (subtractive theory of Table 17.1 with J specified):

$$\hat{\mathbf{R}}_{ij} = a_{\mathbf{R}} \exp(\Psi_{ij}) + b_{\mathbf{R}} \tag{C.1}$$

$$\hat{\mathbf{D}}_{ij} = a_{\mathbf{D}}(\Psi_{ij}) + b_{\mathbf{D}} \tag{C.2}$$

$$\hat{\mathbf{RR}}_{ijkl} = a_{\mathbf{RR}} \exp(\Psi_{ij} - \Psi_{kl}) + b_{\mathbf{RR}} \tag{C.3}$$

$$\hat{\mathbf{RD}}_{ijkl} = a_{\mathbf{RD}}(\Psi_{ij}/\Psi_{kl}) + b_{\mathbf{RD}} \tag{C.4}$$

$$\hat{\mathbf{D}}\mathbf{R}_{ijkl} = a_{\mathbf{DR}}\,(\Psi_{ij} - \Psi_{kl}) + b_{\mathbf{DR}} \tag{C.5}$$

$$\hat{\mathbf{D}}\,\mathbf{D}_{ijkl} = a_{\mathbf{DD}}\,(\Psi_{ij} - \Psi_{kl}) + \mathrm{b}_{\mathbf{DD}} \tag{C.6}$$

$$\Psi_{ij} = s_{A_j} - s_{B_i} \text{ and } \Psi_{kl} = s_{C_k} - s_{D_l} \tag{C.7}$$

Note that the unit of the Ψ values is determined by Eq. C.1. The same unit was assumed for Eq. C.3, as it was expected that judges would be consistent with their previous "ratio" judgments.

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

$$P_{\mathrm{T}} = \frac{\Sigma(X_{\mathrm{T}} - \hat{X}_{\mathrm{T}})^2}{\Sigma(X_{\mathrm{T}} - \bar{X}_{\mathrm{T}})^2} \tag{C.8}$$

where P_{T} is the proportion unexplained, X_{T} is the cell mean judgment, $\hat{X}_{\mathrm{T}}$ the prediction (from Eqs. C.1 through C.7), and $\bar{X}_{\mathrm{T}}$ the mean judgment for task **T** (over all cells). The summation is over all cells in the design for Task **T**. For the "ratio" and "ratio of ratios" tasks, X_{T} is the log cell mean response, $\hat{X}_{\mathrm{T}}$ is the log of the predicted response, and $\bar{X}$ is the mean log response.

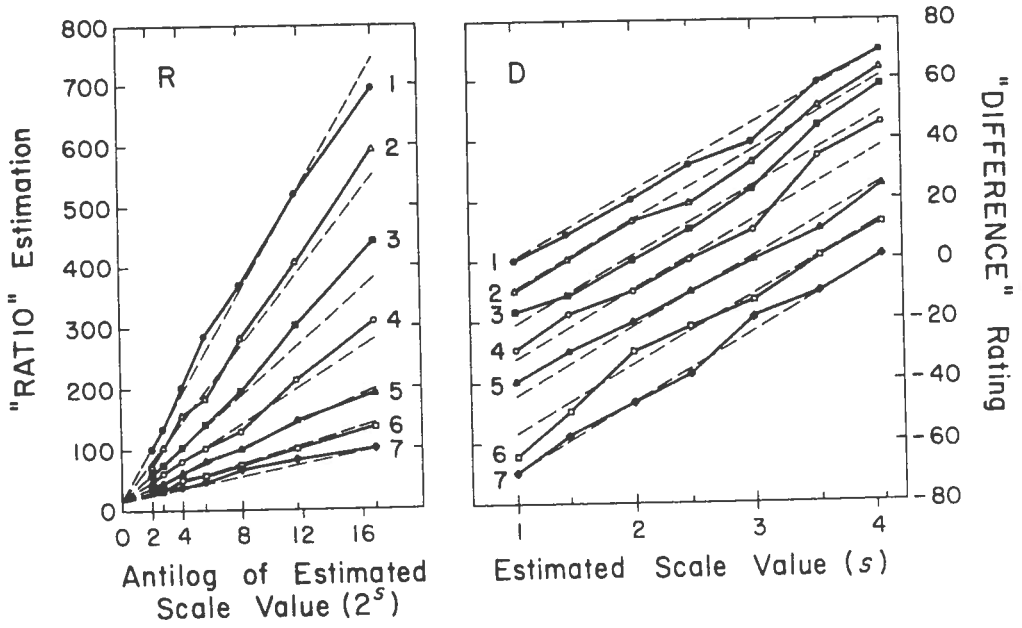


FIG. 17.11. Judgments of "ratios" and "differences" of darkness of dots. Dashed lines are based on the subtractive theory, fit to data of all six tasks simultaneously. Note that "ratios" are plotted against 2^s rather than s, whereas "differences" are plotted against s. Therefore, these data are *not* like predictions of Fig. 17.8, but instead are consistent with subtractive theory (dashed lines). From Birnbaum (1980d).

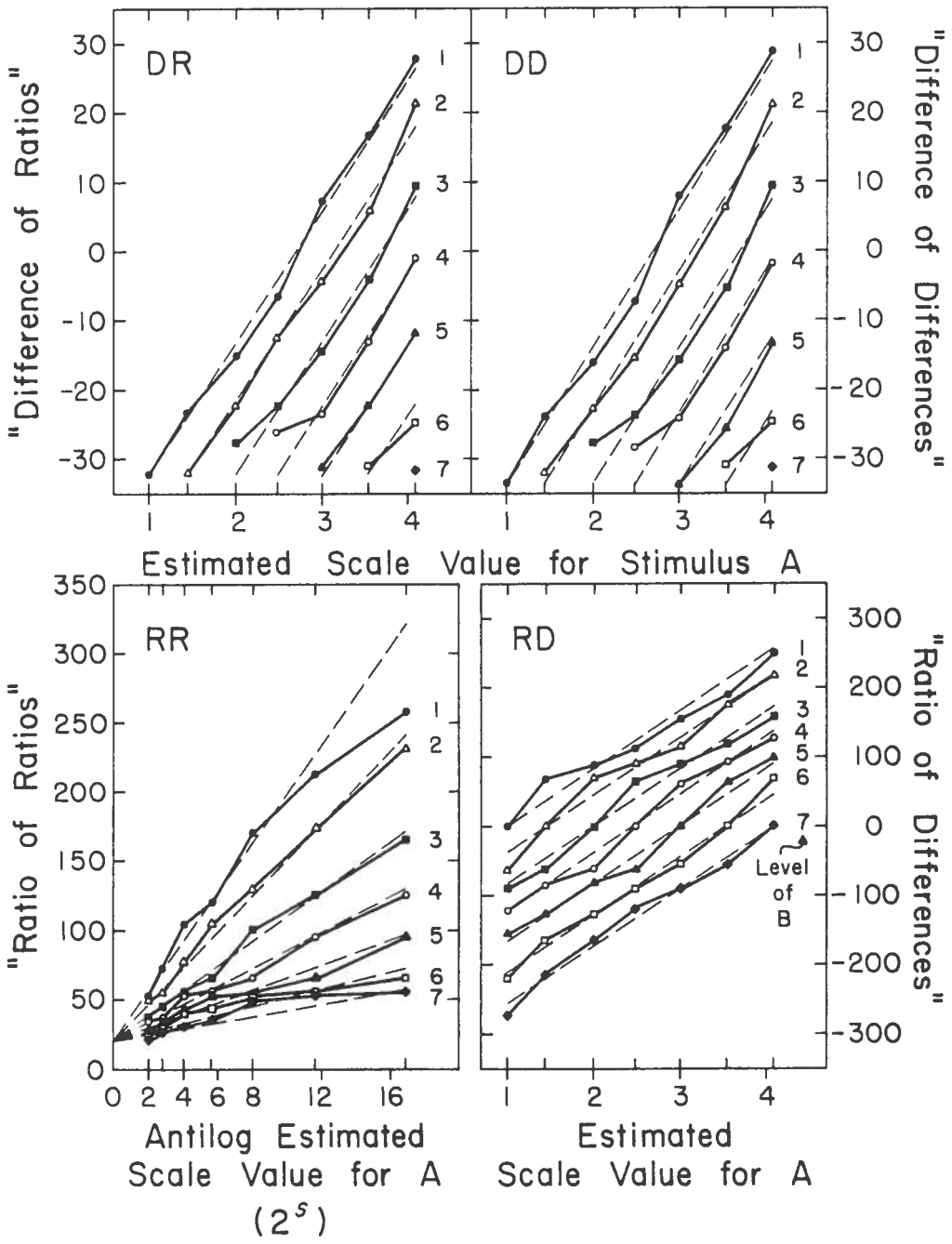


FIG. 17.12. Obtained A-by-B effects, averaged over C and D, for four-stimulus tasks, plotted for comparison with Fig. 17.9. Dashed lines are predictions of subtractive theory, simultaneously fit to all six tasks. Note that abscissa for **RR** task is 2^s rather than s. From Birnbaum (1980d).

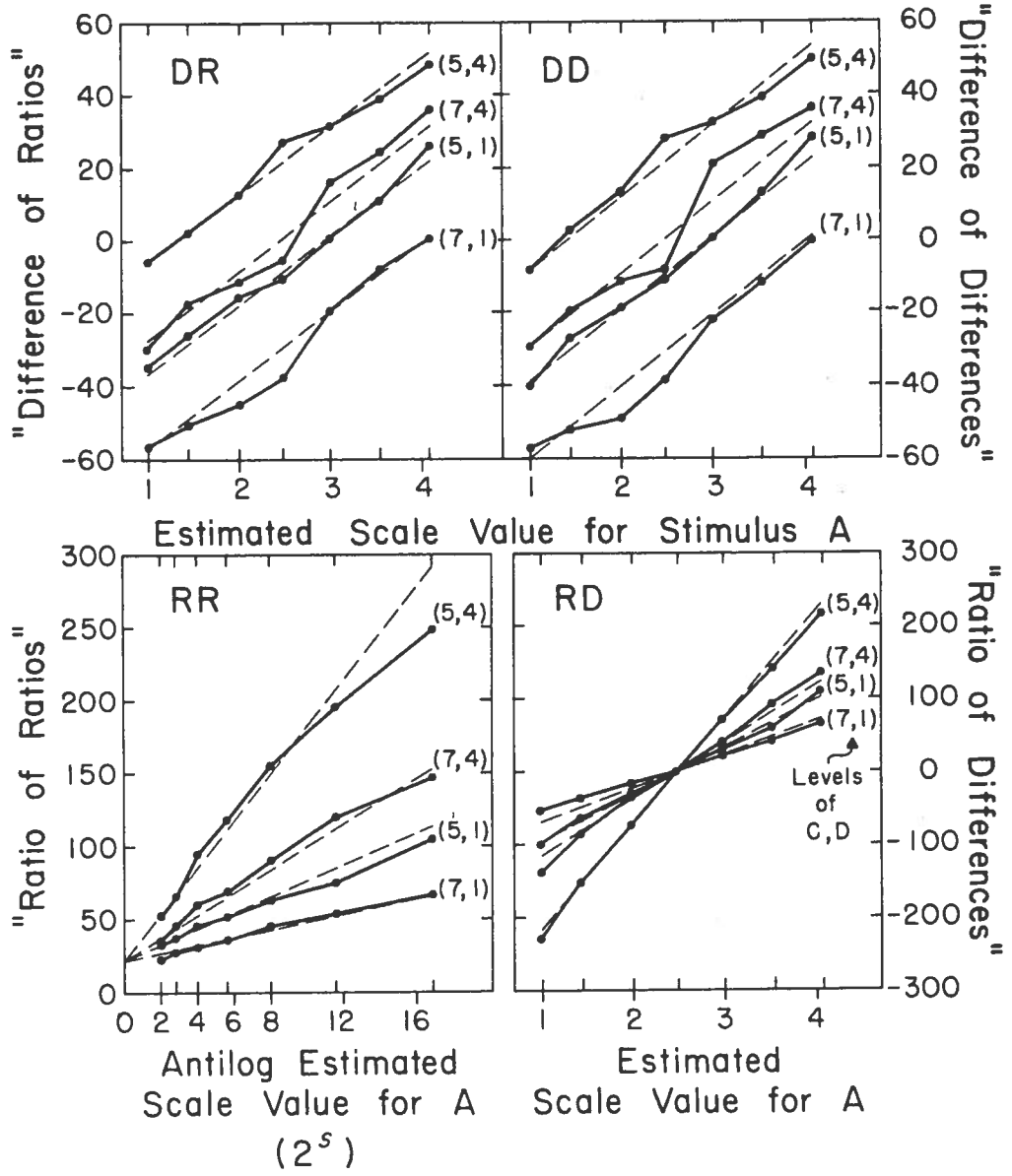


FIG. 17.13. Obtained A-by-C-by-D effects for four-stimulus tasks, plotted for comparison with Fig. 17.10. For **DR** and **DD** tasks, data are for first level of B only. For **RR** and **RD** tasks, data are averaged over levels of B. Dashed lines are predictions of subtractive theory, constrained to fit all six sets of data simultaneously. Abscissa for **RR** task is 2^s. From Birnbaum (1980d).

An overall index, L, was defined as follows:

$$L = \sum_{T=1}^{6} P_T,$$

where L is an index of badness of fit that is the sum over all six data arrays. The **R** and **D** arrays are each 7 × 7, symmetric A by B factorials; the **RR**, **RD**, **DR** and

DD arrays are each 7 × 7 × 2 × 2, A by B by C by D factorials. Therefore, there are 882 cells to be fit. In addition to the 12 linear constants in Eqs. C.1 through C.7, there were six scale values for A and B to estimate (the scale values for s_{A_1} and s_{D_1} are fixed to 1.0) and three scale values to estimate for C and D. Therefore, there are 21 parameters to be estimated from 882 data cells.

The same scale values are assumed for A (and B) throughout all six tasks. The estimated values for the seven levels of A (and B) are 1.00, 1.45, 2.00, 2.47, 3.00, 3.54, and 4.08. The estimated values for C are 3.83 and 4.90, and for D they are 1.0 and 2.58.

The values of P_T for the six tasks for the subtractive theory were .012, .014, .027, .026, .023, and .042 for the **D**, **R**, **DD**, **DR**, **RD**, and **RR** tasks, respectively. Attempts to fit the other theories in Table 17.1 led to poorer overall fits. For example, the sum of squared deviations for the **DR** task for the ratio theory and two-worlds theory was more than 2.7 times as great as for the subtractive theory, with the fit of the other tasks about the same or worse. The subtractive theory gives a reasonable account of the data in Figs. 17.11, 17.12, and 17.13 as shown by the similarity of the solid curves (data) and dashed lines (predictions). The largest deviations, for the **RR** task, may be due in part to the unnecessarily strict restriction that the unit of the exponential in Eqs. C.1 and C.3 were the same. A better fit was obtained by the following: $a_{\mathbf{RR}}\exp[c_{\mathbf{RR}}(\Psi_{ij} - \Psi_{kl})] + b_{\mathbf{RR}}$, where $c_{\mathbf{RR}}$ is less than 1.0.

The subtractive theory assumes that "ratios of differences" can be represented by a ratio of differences model even though "ratios" and "ratio of ratios" are represented by subtraction (Birnbaum, 1978; Hagerty & Birnbaum, 1978; Veit, 1978). Consistent with this assumption, and with the corresponding assumptions concerning J in Eqs. C.1 through C.6, the **RD** data are nearly linearly related to scale values estimated from the other tasks, whereas the **R** and **RR** data are nearly exponentially related.

The other theories make distinct predictions that are not fulfilled by the data. For example, the ratio theory predicts that "differences between ratios" would fit a difference of ratios model. The data in Figs. 17.12 and 17.13 show that the **DD** and **DR** data are nearly identical and do not resemble the predictions of the difference of ratios model (Figs. 17.9 and 17.10).

The indeterminacy theory predicts that the **RD** data could be represented by subtraction. Instead, Figs. 17.12 and 17.13 show the appropriate pattern of parallelism for the A by B interaction and a cross-over interaction for the A by (C − D) interaction. Eisler's (1978) transformation theories (which are discussed in more detail in Section G) predict that **DD** and **DR** should be different and that **DD** should resemble the **RD** data, fitting a ratio of differences model.

In sum, the data of six tasks appear consistent with the pattern predicted by the subtractive theory. This result agrees with the conclusions of Birnbaum (1978, 1979), Veit (1978), and Hagerty and Birnbaum (1978). The results are consistent with the proposition that two stimuli are compared by subtraction whether the

instructions are to report a "ratio" or a "difference." To argue that "ratio" judgments are represented by the ratio model appears to require a complex sequence of arguments (Birnbaum, 1978, 1979).

More Evidence

Rose and Birnbaum (1975) asked judges to divide a line segment to represent either the "ratio" or "difference" of two numbers. The pattern of responses was largely independent of the task. If it was assumed that the responses were represented by a ratio rule:

$$\Psi_{ij} = \frac{s_i}{s_i + s_j} \tag{C.9}$$

then the scale values, s, were found to be a positively accelerated function of numerical value. On the other hand, if the subtractive model was assumed, the scale values were found to be a negatively accelerated function of physical number. The scale values for number estimated from the subtractive model were approximately a linear function of scale values estimated from range-frequency theory (Birnbaum, 1974c) and scale values estimated by other procedures (Rule & Curtis, 1973).

Elmasian and Birnbaum (1979) found that the subtractive theory applied to judgments of "ratios" and "differences" of pitch led to scale values that were compatible with the musical scale, whereas the ratio theory led to scale values that were nonlinearly related to the musical scale of pitch.

Birnbaum and Mellers (1978) asked judges to estimate "ratios" and "differences" of easterliness and westerliness of U.S. cities. The task is a particular "inverse" judgment for which the *inverse* appears an unattractive theoretical interpretation. They used a factorial design that permits segregation of scale values from the response function for "inverse" judgments. As shown in Fig. 17.5, "ratios" were nearly exponentially related to "differences" for both easterliness and westerliness, consistent with the hypothesis that only one operation is involved for both tasks. The data can be well-described by the model:

$$\mathbf{DE}_{ij} = s_j - s_i \tag{C.10}$$

$$\mathbf{DW}_{ij} = s_i - s_j \tag{C.11}$$

$$\mathbf{RE}_{ij} = \exp[a(s_j - s_i)] \tag{C.12}$$

$$\mathbf{RW}_{ij} = \exp[a(s_i - s_j)] \tag{C.13}$$

where $\mathbf{DE}_{ij}$ and $\mathbf{DW}_{ij}$ are the predicted ratings of "differences" in easterliness and westerliness, and $\mathbf{RE}_{ij}$ and $\mathbf{RW}_{ij}$ are the predicted magnitude estimations of "ratios" of easterliness and westerliness, respectively. This theory requires only one cognitive map and one comparison operation (subtraction). Note also that the

distinction between easterliness and westerliness is merely one of direction. Because $\exp(s_i - s_j) = 1/\exp(s_j - s_i)$, it follows that magnitude estimations of "ratios" of easterliness and westerliness are reciprocally related.

Figure 17.14 shows a summary of "mental maps" (scale values) derived from the data of Birnbaum and Mellers (1978). The figure shows that mental maps based on the ratio model depend on direction of judgment and are nonlinearly distorted relative to the actual map. The subtractive model (Eqs. C.10 through C.13) is preferred because it produces a single map for all four tasks that is independent of direction and resembles the actual map closely.

In summary, the ratio theory does not give an adequate account of the four-stimulus results, it leads to an unattractive psychophysical function for number, it contradicts the musical scale of pitch, and it yields mental maps that depend on

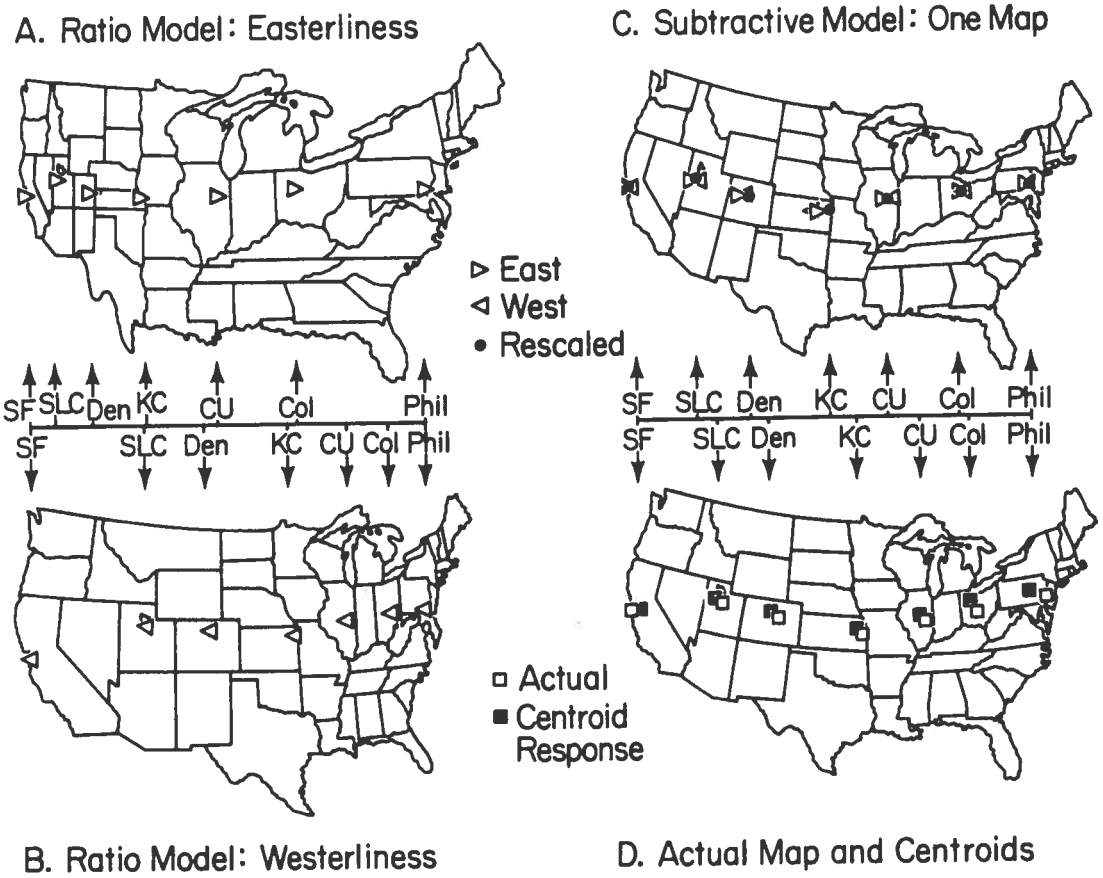


FIG. 17.14. Mental maps of the U.S. derived from ratio model applied to "ratios" of easterliness (Panel A) and "ratios" of westerliness (Panel B). The subtractive model can represent all four sets of data (including "differences") by means of a single map (Panel C). Actual map of U.S. is shown in Panel D. Subtractive theory is preferred over ratio theory because it uses a single map that resembles the actual map, whereas ratio theory requires different maps for different directions that are reciprocally related to each other and nonlinearly related to the actual map. From Birnbaum and Mellers (1978).

the direction of judgment and are nonlinearly related to the actual map. The subtractive theory gives a far simpler summary of these data.

D. CONTEXTUAL EFFECTS IN COMPARISON AND COMBINATION

There are several possible loci for contextual effects in stimulus comparison and combination (Birnbaum, Parducci, & Gifford, 1971) that can be discussed in terms of Fig. 17.7. Contextual effects could operate on the scale values prior to comparison or combination (i.e., on the function H in Fig. 17.7). They could operate *within* the set of stimuli presented for judgment (i.e., on C in Fig. 17.7). Contextual effects could operate on the *between-set* distribution of Ψ or δ in Fig. 17.7 (on the J function). Birnbaum and Mellers (1980b) and Mellers and Birnbaum (1980a, 1980b) have investigated possible loci of contextual effects and reached the following tentative conclusions:

1. When stimuli of the same modality are compared, the distribution of stimulus levels seems to have minimal effects on the scale values.
2. When stimuli from different modalities are compared or combined, the distribution of stimulus levels has large effects on the scale values in the general directions predicted by range-frequency theory.
3. When the context between sets was manipulated in a social judgment task, results were similar to those of Birnbaum et al. (1971, Exp. 5): The contextual effects could be attributed to changes in the judgment function (J).

Context Effects in Within-Mode Comparison

As shown in Fig. 17.3, the judged "ratio" of darkness of stimulus No. 7 (Fig. 17.1, 90 dots) relative to that of stimulus No. 9 (12 dots) can receive a mean judgment anywhere from 3.4, if the largest example is "3", to 8.0 if the largest example is "9". Although this effect was attributed to the judgment function, which describes the relationship between subjective comparisons and overt "ratio" responses, experiments in which the stimuli are varied in one factor (as in Fig. 17.3) do not permit unambiguous identification of the loci of contextual effects.

Given only the results of Fig. 17.3, it is possible that contextual effects operate on the scale values (on H) instead of the judgment function (J). When asked to judge the "difference" or "ratio" of two stimuli, as in Fig. 17.1, will the observer first *judge* each stimulus (with contextual effects) and then *compare* two implicit judgments? Or will contextual effects occur only after stimulus comparison?

These two ideas can be formalized as follows: Let $\mathbf{D}_{ijk}$ and $\mathbf{R}_{ijk}$ be the judged "difference" and "ratio" between stimuli j and i in context k. Suppose J represents the contextual effect (range-frequency). Model 1 states:

$$\mathbf{D}_{ijk} = J_{\mathbf{D}_k}[J_k(s_j) - J_k(s_i)] \tag{D.1}$$

$$\mathbf{R}_{ijk} = J_{\mathbf{R}_k}[J_k(s_j) - J_k(s_i)] \tag{D.2}$$

where $J_{\mathbf{D}_k}$ and $J_{\mathbf{R}_k}$ are the judgment functions (presumably based on the distribution of subjective differences), and J_k is the judgment function for the single stimuli (presumably based on the distribution of s).

Note that if this model holds, then the rank order of "difference" judgments obtained for the positively skewed context (stimuli on the left of Fig. 17.1) would be quite different from the rank order of "differences" for the negative context (on the right). In other words, the estimated scale values $[J_k(s_j)]$ would depend on the stimulus distribution.

Model 2 is a special case of Model 1 that assumes that the scale values are independent of context and that context influences only the transformation from subjective differences to overt judgments. This model can be written:

$$\mathbf{D}_{ijk} = J_{\mathbf{D}_k}[s_j - s_i] \tag{D.3}$$

$$\mathbf{R}_{ijk} = J_{\mathbf{R}_k}[s_j - s_i] \tag{D.4}$$

According to this model, the rank order of "difference" and "ratio" judgments should be independent of stimulus spacing because the scale values are independent of context.

To test these theories, Mellers and Birnbaum (1980b) asked four groups (about 20 undergraduates per group) to judge either "differences" or "ratios" of the darkness of dots spaced in either a positively or negatively skewed context (as in Fig. 17.1). Nested within each 11 × 11 design was a 6 × 6 design of stimuli common to both distributions.

Figure 17.15 shows mean "ratios" plotted against mean "differences" with a separate point for each divisor/subtrahend, plotted as in Fig. 17.4 and Fig. 17.5. Data are shown for the 6 × 6 common design, with the results for the positively skewed condition on the left and the results for the negatively skewed condition on the right. Note that the data appear reasonably consistent with the premise that one operation underlies both tasks, i.e., that "ratios" are approximately a monotonic function of "differences."

Accordingly, the two data sets for each context were fit to the subtractive model:

$$\hat{\mathbf{D}}_{ijk} = a_{\mathbf{D}_k}[s_{jk} - s_{ik}] + b_{\mathbf{D}_k} \tag{D.5}$$

$$\hat{\mathbf{R}}_{ijk} = a_{\mathbf{R}_k}\exp[s_{jk} - s_{ik}] + b_{\mathbf{R}_k} \tag{D.6}$$

where the k subscript on the scale values indicates that different scale values are permitted for each context (though the same scale values and comparison process is assumed for both "ratio" and "difference" tasks). The proportions of variance unaccounted for were computed as in Eq. C.8, and the sum of these proportions was minimized. Parameter estimates were derived from the entire 11 × 11 design in each case. Similar results were obtained when only the 6 × 6 common stimuli were used to fit the model. For the common design, the overall indexes (as in Eq. C.8) were .011 and .014 for positively and negatively skewed conditions, respectively, indicating that the model deviations constitute about half of 1% of the variance for each of the four matrices.

Figure 17.16 shows estimated scale values from two sets of both 11 × 11 matrices. The solid points fall nearly on the identity line, indicating minimal contextual effects. The broken line shows the predicted relationship based on the single judgments for the 100-point scale (Fig. 17.2). The scale values shown in

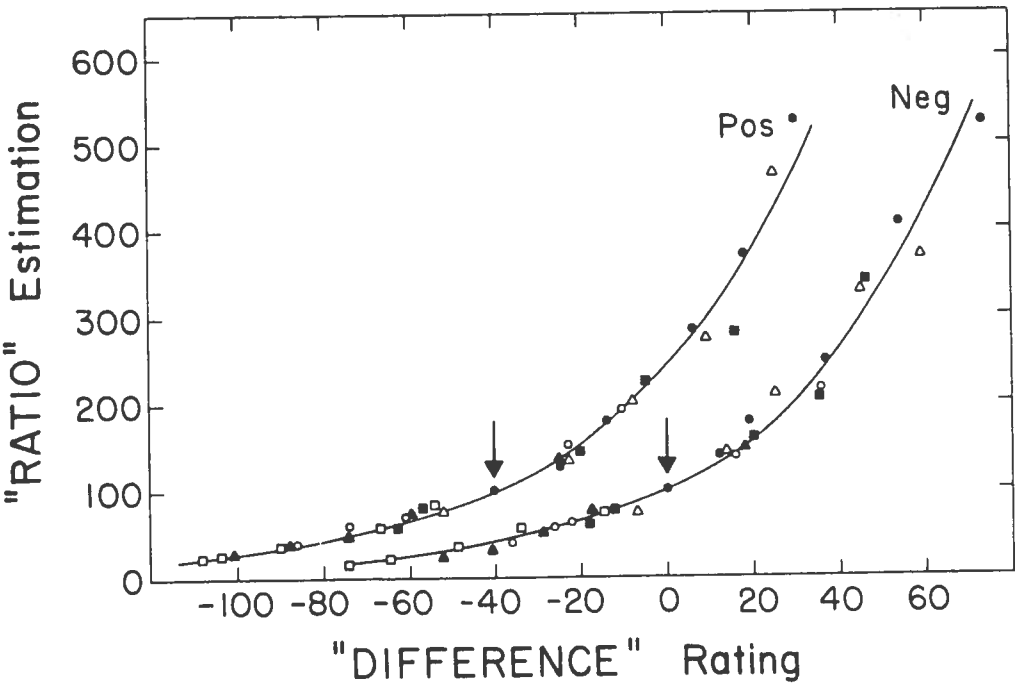


FIG. 17.15. Estimations of "ratios" plotted against estimations of "differences," as in Fig. 17.5. Data on the left are for the positively skewed context (see Fig. 17.1). Positively skewed context data are shifted 40 units to the left relative to the abscissa labels. Curves are best-fit solutions to a special case of the one-operation theory. From Mellers and Birnbaum (1980b).

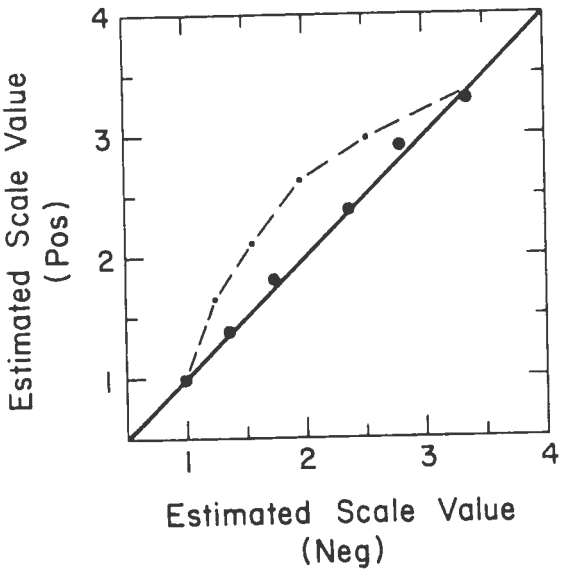


FIG. 17.16. Scale values derived from subtractive theory (applied to both tasks) are largely independent of context. Large solid points show estimated scale values for positively skewed context plotted against estimated scale values for negatively skewed context. Broken curve shows relationship predicted from the 100-point ratings in Fig. 17.2. From Mellers and Birnbaum (1980b).

Fig. 17.16 do not differ as much as would be expected from Model 1, assuming the expected range-frequency compromise. Instead, it appears that one can give a good approximation to the data by the simpler theory (Eqs. D.3 and D.4) that scale values for the subtractive model (applied to both "differences" and "ratios") are independent of stimulus spacing.

Figure 17.2 shows that for the positively skewed condition the difference in judgment between stimulus 90 and 40 is *less than* the difference in judgment between 27 and 12, whereas for the negatively skewed distribution the order of these differences in judgment is reversed. However, when judging "differences" (or "ratios"), the judged "difference" (and "ratio") between 90 and 40 *exceeds* the "difference" (and "ratio") between 27 and 12 for all four comparisons of means for the negative skew and three of four for the positive skew.

Thus, these data do not provide evidence that contextual effects operate on s in within-mode stimulus comparisons. Instead, it appears that the rank order of "ratios" and "differences" can be reproduced by assuming that judges compute differences between scale values that are independent of the stimulus spacing.

Cross-Modality Combination and Comparison

Aside from the psychophysics laboratory, cross-modality questions such as "Who was the greater, Babe Ruth or Roman Gabriel?," "Does the punishment fit the crime?," "Is this salary fair for this job?" are often asked. It seems likely that responses to such questions would not depend solely on absolute values but would depend on the joint distribution of the two modalities.

Krantz (1972) discussed mapping and relation theories of cross-modality "matching." According to the mapping view, sensations in different modalities are somehow mapped into a common scale of magnitudes that can be compared.

According to the relation view of Shepard (1978) and Krantz (1972), relationships (e.g., ratios) between pairs of stimuli can be compared. In other words, it is possible to compare the ratio of two heavinesses to the ratio of two loudnesses. By analogy with physical measurement (in which lengths cannot be compared with masses, but ratios of lengths can be compared with ratios of masses), the relation theory seems sensible.

Another view can be called psychological relativity theory (Birnbaum & Mellers, 1980b). In this theory, each stimulus is compared to its distribution, and the relative positions of the two stimuli in the two modalities are compared. Thus, a loudness will be "matched" to a brightness when the two stimuli hold the same position in the distributions of their respective modalities.

To study possible dependence of the scale values on the stimulus distribution, Birnbaum and Mellers (1980b) investigated two tasks: cross-modality "difference" judgments and "total" intensity judgments. "Total" intensity judgments have been studied by Feldman and Baird (1971) and Anderson (1974a).

A typical stimulus presentation is shown in Fig. 17.17. The "difference" task was to compare the size of the circle to the darkness of the dot pattern and judge which is greater and by how much. The "total" task was to combine the size of the circle and the darkness of the dot pattern. On some trials, only one stimulus (dot pattern or circle) was presented. On these occasions, the unpresented stimulus value was assumed to be zero.

The experimental design paired each of six circles, varying from 8 to 25 mm in diameter factorially with six common dot patterns geometrically spaced from 12 to 90 dots. In two conditions, the positively and negatively skewed distributions (of Fig. 17.1) were factorially combined with the six circles. In two other conditions, more extreme patterns of 10 and 135 dots were added for the medium range; or patterns of 6 and 180 dots were added for the wide range.

The models were as follows:

$$\mathbf{D}_{ijk} = a_{\mathbf{D}_k}(c_j - d_{ik}) \tag{D.7}$$

$$\mathbf{T}_{ijk} = a_{\mathbf{T}_k}(c_j + d_{ik}) + b_{\mathbf{T}_k} \tag{D.8}$$

where $\mathbf{D}_{ijk}$ and $\mathbf{T}_{ijk}$ are the judgments of "difference" and "total intensity" of circle j and dot pattern i in context k, c_j and d_{ik} are the scale values of the circles and the dot patterns, respectively, and $a_{\mathbf{D}_k}$, $a_{\mathbf{T}_k}$, and $b_{\mathbf{T}_k}$ are linear constants for each context.

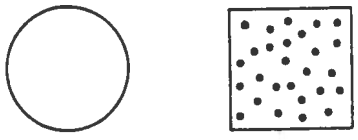


FIG. 17.17. Example stimulus array for one trial of cross-modality comparison. From Birnbaum and Mellers (1980b).

In the subtractive model, when the response is "no difference" it is assumed that $c_j = d_{ik}$, i.e., a cross-modality "match." In the additive model, the additive constant, $b_{\mathbf{T}}^{k}$, is determined by the constraint that when a stimulus is not presented its value is zero. Therefore, $\mathbf{T}_{iOk} + \mathbf{T}_{Ojk} - \mathbf{T}_{ijk} = b_{\mathbf{T}}^{k}$, where $\mathbf{T}_{iOk}$ and $\mathbf{T}_{Ojk}$ are judgments of the ith dot pattern alone and jth circle alone in context k.

Scale values for the dot patterns were estimated separately for "totals" and "differences" for each context. Inasmuch as the distribution of circles was the same for all conditions, circle scale values were assumed to be the same for all conditions. Because the scale values for the circles are assumed to be the same across conditions, once the unit of the circle scale values is fixed, the estimated scale values for dot patterns in the different contexts are uniquely determined for the "total" task. Similarly, once the unit and additive constant for the scale values of circles is fixed, the scale values of the dot patterns for the "difference" task are uniquely determined by the data.

Estimated scale values are shown in Fig. 17.18 as a function of log Φ with separate curves for each context. Note that for both "differences" and "totals," the slopes are greater for the narrow-range conditions (positive and negative skew) than for the medium- or wide-range conditions. The wide-range condition was the lowest in slope. Note also that the scale value of a medium-level dot pattern (e.g., 27 or 40) receives a greater scale value in the positively skewed context than it does in the negatively skewed context. These contextual effects, which are in the general direction of the usual contextual effects in ratings, cannot be attributed to the J functions between Ψ and response, for different circles are judged to "match" the same dot patterns in different contexts.

In previous tests of cross-modality "matching," experiments have controlled the stimulus distributions to be comparable in the magnitude estimation and cross-modality "matching" experiments. It seems likely that if the stimulus and response ranges were systematically manipulated, then the conclusions of cross-modality matching experiments would be altered. Let $\mathbf{\Delta R}$ be the log response range and $\mathbf{\Delta S}$ be the log stimulus range, then the power-function exponent in a magnitude-estimation experiment will be $b = \mathbf{\Delta R}/\mathbf{\Delta S}$. If $\mathbf{\Delta R}$ is a constant (Teghtsoonian, 1971), then exponents for two modalities will be $b_1 = \mathbf{\Delta R}/\mathbf{\Delta S}_1$ and $b_2 = \mathbf{\Delta R}/\mathbf{\Delta S}_2$. Thus, the predicted cross-modality exponent is $b_1/b_2 = (\mathbf{\Delta R}/\mathbf{\Delta S}_1)/(\mathbf{\Delta R}/\mathbf{\Delta S}_2) = \mathbf{\Delta S}_2/\mathbf{\Delta S}_1$. In other words, one should be able to predict cross-modality exponents either from the stimulus ranges or from the magnitude-estimation exponents.

To unconfound these different interpretations, cross-modality "matching" experiments should systematically manipulate the stimulus and response ranges. For force of handgrip, the response range is not under the experimenter's control but rather under the subject's control. A better procedure would be to use a dimension such as loudness for the response and to vary the range and taper of the control knob. It seems likely that the cross-modality "matching" function will depend heavily on the range and distribution of responses under the subject's control. As

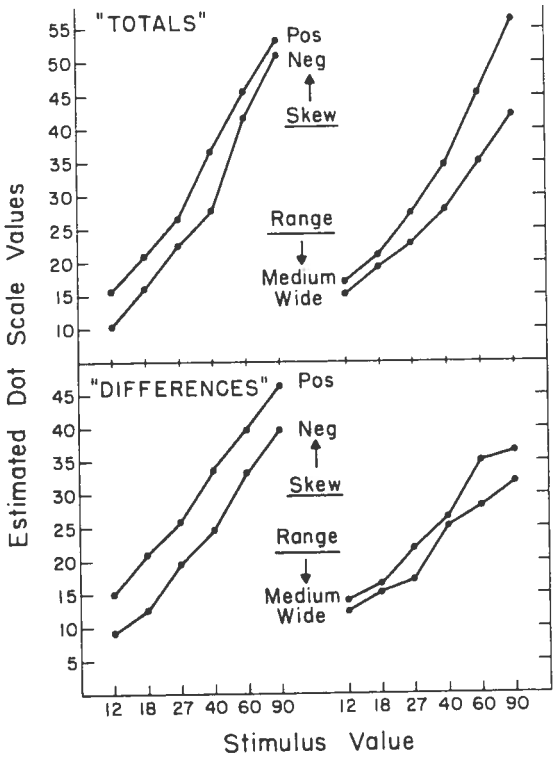


FIG. 17.18. Estimated scale values for dot patterns from studies of cross-modality comparison and combination. If there were no contextual effects, curves should all coincide. Differences in height and slope of curves are generally in the direction predicted by range-frequency theory. From Birnbaum and Mellers (1980b).

another example, the response could be lifted weight. The experimenter could present a number of bottles with the judge's task being to select the bottle whose heaviness "matches" the stimulus. The distribution of weights is clearly under the experimenter's control.

In summary, experiments show that in cross-modality comparison or combination, the range and spacing of the stimuli affect the scale values. It is as if stimuli must be judged within the context of other stimuli in the same modality before they can be compared across modalities. This result is compatible with a relativity theory of cross-modality "matching" rather than the mapping or relation views.

Contextual Effects in Social Information Integration

Mellers and Birnbaum (1980a) applied the approach of Birnbaum et al. (1971, Experiment 5) to a social judgment task in which judges evaluated the performance of students on the basis of their scores on two exams.

The joint distribution of exam scores for the positively skewed context is shown in Fig. 17.19. Each symbol represents the exam scores of one or more hypothetical students. The solid squares show the common stimuli that were presented in both contexts. The common stimuli consist of the union of 4 by 7 and 7 by 4, first exam by second exam, factorial designs. Each open circle represents a contextual trial; each open triangle represents three such trials. Thus,

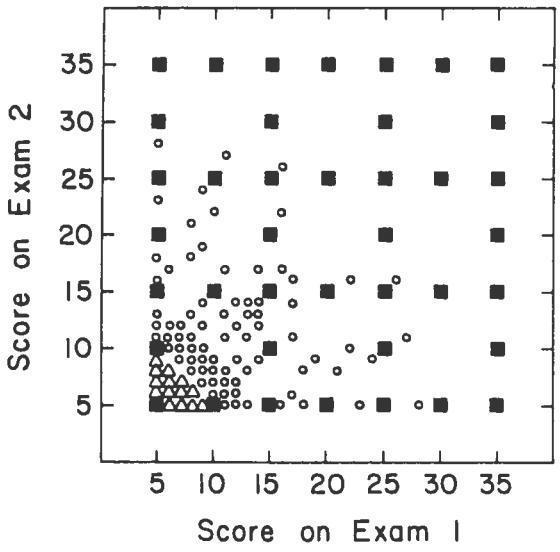


FIG. 17.19. Joint distribution of exam scores for experiment on contextual effects in social-information integration. Figure shows distribution for positively skewed condition. Each open circle or solid square represents the performance of one hypothetical student. Each triangle represents three students with the same scores. Solid squares were the same for both contexts. Contextual trials for negatively skewed distribution were the mirror image of those in this figure, reflected about the axis, exam 1 + exam 2 = 40. From Mellers and Birnbaum (1980a).

the total of both exam scores ranges from 10 (i.e., 5 + 5) to 70 (i.e., 35 + 35). The distribution of total score and the marginal distributions of each exam score are positively skewed. The open triangles show that 42 students (of the total of 160) had scored scores between 10 and 15 for the positively skewed distribution. There were none in this open interval for the negatively skewed context, which had a mirror-image distribution, reflected about the axis, exam 1 + exam 2 = 40.

Assuming the scores are combined by an additive (or parallel-averaging) model, the general model of context can be written:

$$\mathbf{G}_{ijk} = J_k^*[J_k(s_i) + J_k(s_j)] \tag{D.9}$$

where $\mathbf{G}_{ijk}$ is the evaluation of the student with scores i and j in context k, J_k^* is the judgment function (presumably based on the distribution of Ψ_{ij}), and J_k represents the contextual effect on the scale values. In general, if J_k is nonlinear, then the rank orders for different contexts will be different.

A variation of the model that assumes parameter invariance but does not assume additivity can be written

$$\mathbf{G}_{ijk} = J_k^*[\Psi_{ij}] \tag{D.10}$$

where Ψ_{ij} is the integrated impression, which may or may not be additive. This model implies that the rank order of the data should be independent of context.

There were four groups. Half of the judges received either the positively or negatively skewed distributions. Half of each of these groups were given histograms depicting the marginal distributions of exam 1 and exam 2 to use while making their judgments. It was thought that presentation of these histograms might enhance any tendency to evaluate performance first and then combine, as in Eq. D.9. There were about 25 subjects in each of the four groups.

The mean judgments are plotted in Fig. 17.20 as a function of the score on exam 1 with a separate curve for each level of exam 2. Parallelism would be

consistent with an additive (or parallel-averaging) model. Instead of being parallel, however, the curves for the positively skewed context show systematic convergence to the right (the vertical separations between the curves decrease as the score on exam 1 increases). The curves for the negatively skewed context show the opposite: divergence to the right. Thus, the apparent interaction between the exam scores depends on the stimulus distribution. This result is consistent with the results of Birnbaum et al. (1971), who used psychophysical stimuli.

The interaction can be represented in this case by assuming that only the J^* function (rather than the scale values or combination process) depends on the stimulus distribution, as in Eq. D.10. Note that the rank orders of the data points are essentially the same in all four panels. The rank orders would be systematically different in the different panels had the subjects made separate (context-dependent) judgments of each exam score and averaged their separate judgments.

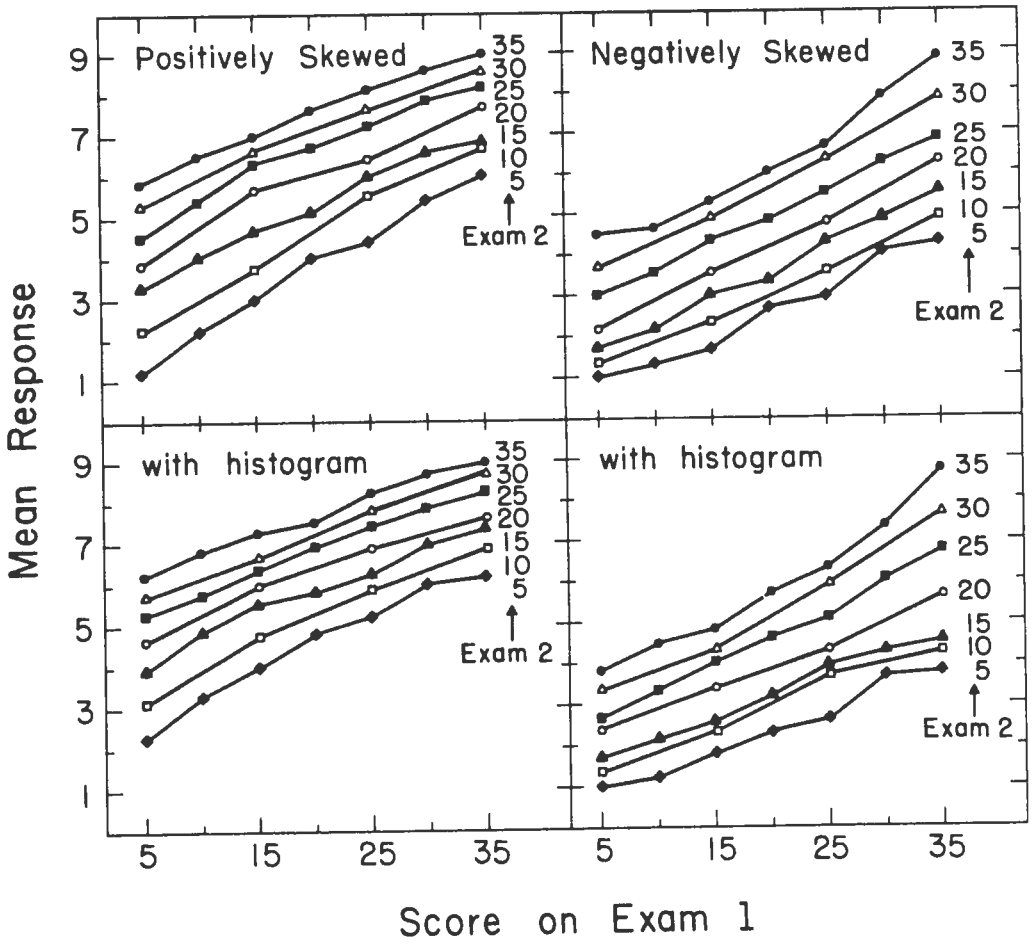


FIG. 17.20. Mean evaluations of the performance of students, as a function of score on exam 1, with a separate curve for each level of score on exam 2. Note that curves converge for positively skewed distribution (left panels), and they diverge for the negatively skewed distribution (right panels). From Mellers and Birnbaum (1980a).

(To check this, another group of subjects was asked to judge performance based on exam 1 only. Indeed, when these judgments were substituted into Eq. D.9, the rank order of predictions differed for different contexts.)

The data of Fig. 17.20 were rescaled to parallelism, and scale values estimated from the additive model applied separately to the four conditions were found to be nearly identical across conditions. These estimates were averaged and used to compute $s_i + s_j$. The mean responses are plotted in Fig. 17.21 as a function of $s_i + s_j$, with a separate curve for each context. (The data were averaged over the histogram vs. no histogram manipulation, which showed minimal effects in Fig. 17.20.) The mean judgments have been linearly calibrated to the same zero to one scale in Fig. 17.21. The solid curves show the cumulative density functions for the two contexts, based on the density of the sum of the exam scores (see Fig. 17.19).

According to range-frequency theory, if $\Psi_{ij} = s_i + s_j$, then the obtained values for each condition should be an average of the solid curve for that context (frequency) and a straight line through the end points (range). The dashed lines show the predictions of this theory.

The dashed curves give a good approximation to the data. It is also possible to use range-frequency theory to solve for the values of Ψ_{ij} in order to determine if the assumption that $\Psi_{ij} = s_i + s_j$ is reasonable. The model was fit as follows:

$$\mathbf{G}_{ijk} = aF_k(\mathbf{G}_{ijk}) + \Psi_{ij} + c \qquad \text{(D.11)}$$

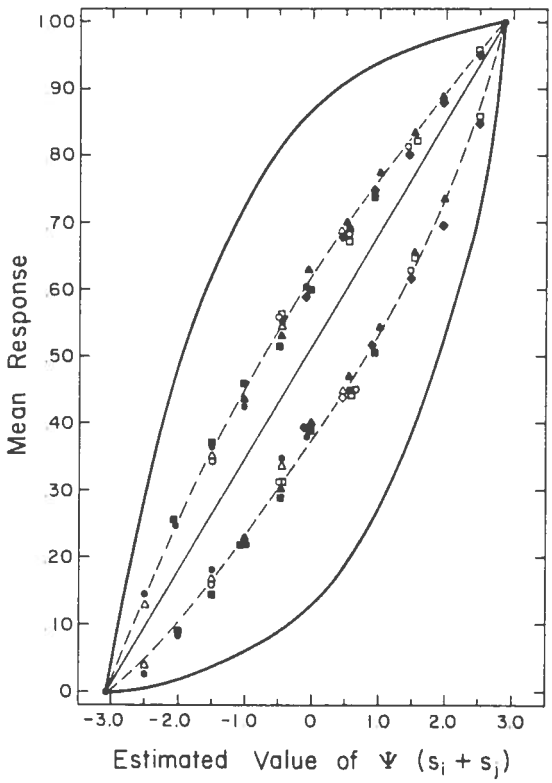


FIG. 17.21. Mean judgments of student performance from Fig. 17.20 (recalibrated to a 0–1 scale) as a function of estimated value of Ψ, with separate dashed curves for each contextual distribution. Solid curves are cumulative density functions for the two contexts, dashed curves are predictions of range-frequency theory. Symbols are consistent with Fig. 17.20. From Mellers and Birnbaum (1980a).

where $\mathbf{G}_{ijk}$ is the rating of a student with test score levels i and j in context k, F_k is the cumulative proportion of students receiving lower judgments (curves in Fig. 17.21), Ψ_{ij} are estimated parameters, and a and c are fitted constants. The model was fit by means of multiple regression using dummy variables. The estimated values of Ψ_{ij} were very nearly parallel (with a very small divergence). This analysis gave only a slightly better fit than the special case that assumed $\Psi_{ij} = s_i + s_j$. The additive model therefore appears to provide a satisfactory approximation.

In summary, Figs. 17.20 and 17.21 show that the data are consistent with the hypothesis that manipulation of the distribution of totals (as in Fig. 17.19) can manipulate the judgment function, according to the principles of range-frequency theory, applied to the distribution of totals. No evidence was found to require the interpretation that the scale values for the test scores depended on the marginal distributions. Indeed, had the scale values changed according to range-frequency theory, the rank order of the points in Fig. 17.20 would have changed.

These results show the potential importance of contextual effects in studies of information integration in which the parallelism test is of interest for evaluating theories. In this case, it should be clear that the change from convergence to divergence was brought about by the change in stimulus distribution and can be explained in terms of a range-frequency analysis of the judgment function. This experiment shows that it should often be possible to select stimuli to produce parallelism, even though the model is not additive. The implications of this finding for functional measurement are further discussed in Section F.

Contextual Effects in Similarity Judgments

Tversky (1977) argued that distance theories of "similarity" judgments cannot account for violations of the three basic axioms of a distance function, δ:

Minimality:	$\delta(x,y) \geqslant \delta(x,x) = 0$
Symmetry:	$\delta(x,y) = \delta(y,x)$
Triangle inequality:	$\delta(x,y) + \delta(y,z) \geqslant \delta(x,z)$

Tversky notes that judgments of the form, "North Korea is like Red China" often violate symmetry because "Red China is like North Korea" is less preferred. He has also questioned the other axioms. Tversky's (1977) theory, which extends developments of Restle (1959), is that judgments of "x is like y" depend on the common psychological features of x and y, the features that x possesses that are absent from y, and those features belonging to y but not x. In Tversky's theory, measures of these sets are permitted to depend on the context, but no theory is advanced to describe the effects of context.

Krumhansl (1978) argued that the distance concept can be saved by introducing a particular theory of contextual effects in "similarity" judgments. Her theory assumes that "similarity" judgments depend on both distance in a

context-free multidimensional space (that satisfies the axioms) and also on local stimulus densities in the regions of the points.

There are many possible loci for contextual effects in the distance model. Investigating them empirically would require a large experimental effort. Figure 17.19 can be relabelled to facilitate discussion of possible loci of contextual effects. Suppose the abscissa represents (psychological) dimension I, and the ordinate represents (psychological) dimension II. There are then three likely loci for contextual effects: (1) projections of points on the dimensions (scale values); (2) distance calculations (Ψ_{ij}); (3) the judgment function relating overt judgments of similarity to subjective distances.

Suppose, for example, the judge's task was to rate the "similarity" between pairs of squares that vary in size and reflectance. There are three kinds of distributions to consider: (1) the marginal distribution of square sizes and the marginal distribution of square darknesses; (2) the joint distribution of sizes and darknesses; (3) the distribution of distances (which depends on the pairs presented for similarity judgment).

Manipulation of the marginal distributions in such a task seems analogous to the cross-modality experiments of Mellers and Birnbaum (1980a). It seems likely that manipulation of the range and spacing of the levels will affect the scale values (projections on the axes). It also seems likely that the difference in size coordinates between two given squares will vary with the total range of squares and the number of squares intermediate in size. Variation in stimulus spacing would presumably produce nonlinear changes in the projections (of the stimuli on the axes) in the usual multidimensional scaling solution.

Variation in the joint distribution would be expected to affect distance judgments so as to increase the judged dissimilarity between two points that have a large number of points "between" them in space. Krumhansl's (1978) model attempts to deal only with this aspect of contextual effects. However, her model does not reduce to range-frequency theory in one dimension, because it deals only with the densities in the regions of the points rather than in the space between them. It may be preferable to define the density term as a weighted integral of the stimulus density within an ellipsoid (based on the two points x and y as foci) where the weights are a function of the distance from x. It seems likely that the judged distance between two points will be greater when the stimulus density in this region "between" points is greater. This version would reduce to range-frequency theory in one dimension with suitable ellipse and weighting function.

E. SYSTEXTUAL DESIGN

There is a fundamental difference between physics and psychophysics that has long troubled psychologists (Baird & Noma, 1978; Luce, 1972). In classical physics, the measuring devices do not "remember" their previous mea-

surements. Measurements of length or mass, for example, do not depend on the other lengths or masses previously measured. However, human judges give different responses to the same stimulus depending on the other stimuli forming the context for judgment, as illustrated in Sections A and D. Obviously, numerical judgments of subjective values cannot be regarded as analogous to the readings of voltmeters or thermometers. Figures 17.2 and 17.3, for example, show that category ratings and magnitude estimations of single stimuli depend on the stimulus spacing and response range. Figure 17.18 shows that scale values derived from cross-modality comparisons and "total" intensity judgments depend on the stimulus distribution. Figure 17.20 shows that the test of parallelism depends on the distribution of combinations.

Because the results of psychological experiments depend on the distribution of treatments to which the subject is exposed, psychologists have become concerned with the implications for generalization. At least four distinct methodological positions have emerged: standardized design, representative design, between-subjects design, and systextual design. In standardized design, the context is fixed to some conventional value. In representative design, the aim is to survey the environment and use the context to which generalization is desired. Between-subjects designs hope to "avoid" the context by allowing each subject to choose his or her own context. Systextual design systematically manipulates context.

Standardized Design

In standardized design, procedures for the conduct of research are agreed upon. If scientists all agree to do the same experiment, they should all obtain the same results. This position assumes that certain variables can cause a nuisance when left uncontrolled by different experimenters. In physics, for example, a calorie has been defined as the amount of heat required to raise one gram of water one degree Celsius at 4°C. Because the heat required to raise water temperature by a given amount depends on the temperature, it is necessary to qualify the temperature, that is, to agree to standardize our measurements. Much can be said in favor of the reasonableness of this approach. If psychology can develop consistent laws that hold in some restricted domain (no matter how restricted), we will have the beginnings of a science.

One unfortunate offshoot of the approach of standardization has been what can be called a *standardization of circularity* (perhaps analogous to the use of persuasive definition in philosophy). The circular standardization argues that the "right" way to do research is by the method that yields results compatible with a pet theory and therefore anyone who deviates from the approved method cannot be taken seriously. For example, it is possible to adjust the spacing of the stimuli so that the data for a magnitude-estimation study actually approximate a power function of physical value. It is also possible to select the stimulus spacing to produce deviations from a power function (as in Fig. 17.3). Because the power

function was once thought to be the "right" function, any procedure that maximized the fit of it was deemed a "good" procedure. As Poulton (1979) has remarked, finding a stimulus spacing that produces a fit to the power function does not provide support for the power function; it merely demonstrates that the researcher has knowledge (at some level, perhaps implicit) of the stimulus spacing effect. Surprisingly, Poulton (1979) advocates geometric spacing as the "right" procedure, though he offers no theoretical justification. Stevens, Anderson, and others have given (conflicting) pronouncements concerning proper procedures for psychophysical studies, and their suggestions have unfortunately become orthodoxy to many persons.

Perhaps a time may come for psychologists to agree to adopt a set of standardized procedures. However, such agreement should not (and hopefully will not) occur until the scientific questions under investigation have been settled. Until that time, there is a danger that a theory will be accepted prematurely and will bring with it a set of orthodox, "right" procedures that will prevent its modification or falsification.

Criteria for Evaluating Methods

A set of criteria for evaluating psychophysical methods would seem useful. The following criteria, which have been implicitly suggested, do *not* seem appropriate ones:

1. Proper methods are those that yield results consistent with the power law. (Problem: The power law may not be appropriate.)
2. Proper methods are those that yield data that are parallel (fit the additive model). (Problem: The model predicting parallelism may not be valid.)
3. Proper methods are those that avoid the possibility of testing invariance properties of the scales. (Problem: By avoiding the possibility of testing the invariance properties, one does not establish invariance, one merely avoids the issue.)

The following considerations seem more useful for evaluating psychophysical methods and theories.

1. A psychophysical scale consists of a set of scale values used to reproduce the rank order of empirical data in terms of a theory.
2. The value of the scale is enhanced if it can be shown that the scale values cannot be arbitrarily transformed (beyond the uniqueness of the model) and still reproduce the rank order of the data. In other words, the more the scale is constrained by the data the better. (One can consider the scale values to be parameters estimated from data.)
3. A psychophysical scale should not only operate in a single situation, but it should show generality across situations (scale convergence). Thus, a theory involving one set of scale values and a pair of theories for two

empirical situations is preferred to one using two different scales and two different theories.

Methods that allow one to assess the scales in terms of these three considerations should be preferred over methods that do not permit these tests.

Consider Poulton's (1979) suggestion to use equal geometric spacing of the stimuli in magnitude-estimation experiments. How can one decide if this suggestion is reasonable? How do we know we are obtaining the "true" result? The experimental design *precludes* the possibility of establishing contextual invariance or demonstrating that the "true" result has been obtained.

Between-Subjects Design

Poulton (1979) has listed many of the factors that influence the outcome of direct scaling studies (as in Figs. 17.2 and 17.3) and has argued that contextual effects (which he regards as "biases") can and should be "avoided." Poulton (1973) suggested that such effects can be "avoided" by using a special type of standardized design in which the observer is presented with only one level of the treatment. A standardized design in which each subject receives only one treatment combination is called a between-subjects design.

The idea behind between-subjects designs (and certain other suggestions by Poulton for experimental procedure) is that there is an ideal laboratory condition that can and should be achieved in psychology. In physics, for example, Galileo's law of falling bodies would not be verified in the atmosphere. Galileo's critics noted that when two objects are dropped, the coin strikes the earth before the feather. In Galileo's time, nature still abhorred the vacuum, so it was not possible to conduct the experiment under reduced atmospheric conditions. But today, many museums demonstrate a low pressure tube in which coin and feather do fall together. Being confident in Galileo's premise, we now feel that the vacuum is the "right" context for conducting the study.

Poulton's (1973, 1979) suggestions to "avoid" contextual effects seem based on the proposition that they are analogous to friction in the physics lab. For example, in response to the stimulus spacing effects and effects of examples in the instructions (as in Fig. 17.3), Poulton (1979) recommends presenting the subject with only one stimulus and giving the subject no examples. Unfortunately, we cannot achieve a psychological vacuum in our judges' minds by presenting them with only one stimulus. Just because we have not presented other stimuli for judgment does not mean that our subjects, who are usually adults, have never before experienced a stimulus. There are two kinds of contexts: the context the subject brings to the laboratory and the context provided in the laboratory. Subjects' judgments depend on both. Therefore, when a subject is given a single stimulus to judge, the subject brings extralaboratory contexts to the task. It is even possible that when a different stimulus is presented to each subject that the context will be confounded with the stimulus.

Confounding of Contexts in Between-Subjects Designs

It may be that many of the counterintuitive findings so well-liked in social psychology are merely results of the *confounding of contexts* that can occur in between-subjects designs. To explain this point, consider an experiment by Jones and Aronson (1973) on the judged fault of rape victims. Which victim is most at fault for her own rape: the housewife, virgin, or divorcée? Jones and Aronson (1973) presented each subject with *only one* case history and found that the divorcée was rated *least* at fault and the virgin and married woman were *more* at fault (see Fig. 17.22). They interpreted this result in terms of a "just world" hypothesis. In a "just world," you get what you deserve. What you deserve (presumably) depends on *who you are* and *what you do*. If the victim did not deserve to be raped because of *who she was* (e.g., respectable, married), she must have *done something* to deserve it, and therefore, according to this theory, the more respectable victim should be rated *more* at fault.

It is difficult, but possible, to replicate the Jones and Aronson (1973) experiment using a between-subject design. In 1973 at Kansas State University, the effect was observed in a between-subject design only if the fault of the defendant was not rated, but the effect was reversed if both victim and defendant were rated (see Fig. 17.23). At the University of Illinois, at a time when a local rapist was causing great concern on campus, in a between-subjects design it was found that 10.5%, 10.4%, and 4.7% of 76, 67, and 85 subjects thought the virgin, housewife, and divorcée, respectively, were at fault exceeding 15 (on a 1–20 scale) for their own rape.

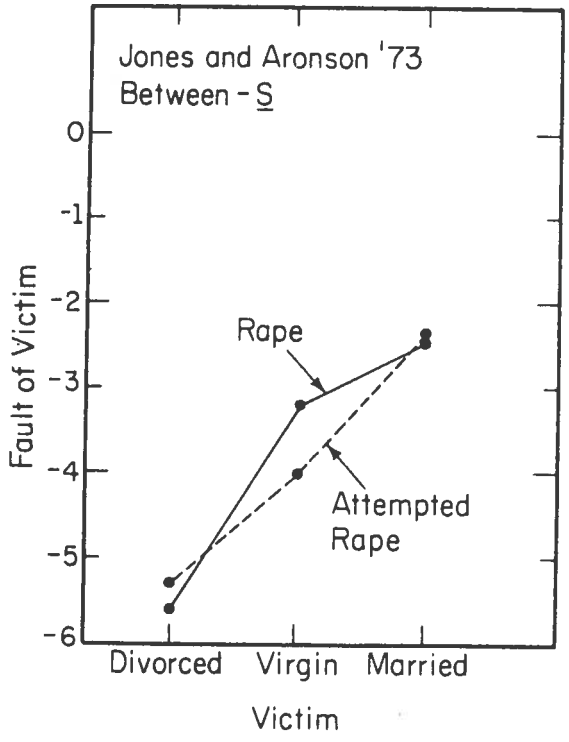


FIG. 17.22. Judged fault of rape victim as a function of victim's respectability in a between-subject design (Jones and Aronson, 1973). One group of subjects rated the fault of the divorced victim to be less than the average rating of fault given by another group of subjects who rated only the raped virgin. From Birnbaum (1980c).

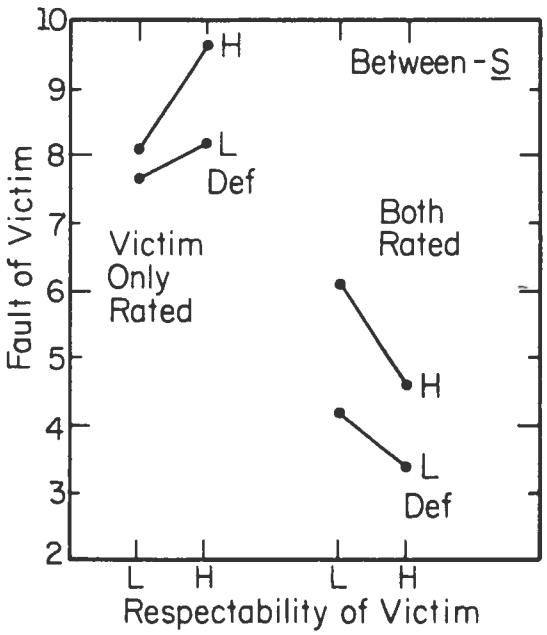


FIG. 17.23. Mean judgments of fault of victim as a function of victim's respectability, with a separate curve for each level of defendant's respectability. Each point is based on the mean judgment of a different group of judges who received only one type of case history. Data on left are for judges who rated only the victim. Data on the right are for judges who rated the fault of both victim and defendant. From Birnbaum (1980c).

However, in a within-subject design (Birnbaum, 1980c), it has been found that the judged fault of a victim *decreases* with increasing victim respectability. Judged fault of the victim is also greater when the defendant is higher in respectability and is lower for more severe crimes. These results are shown in Fig. 17.24, which plots judged fault as a function of the respectability of the victim, with a separate curve for each level of respectability of the defendant. It there-

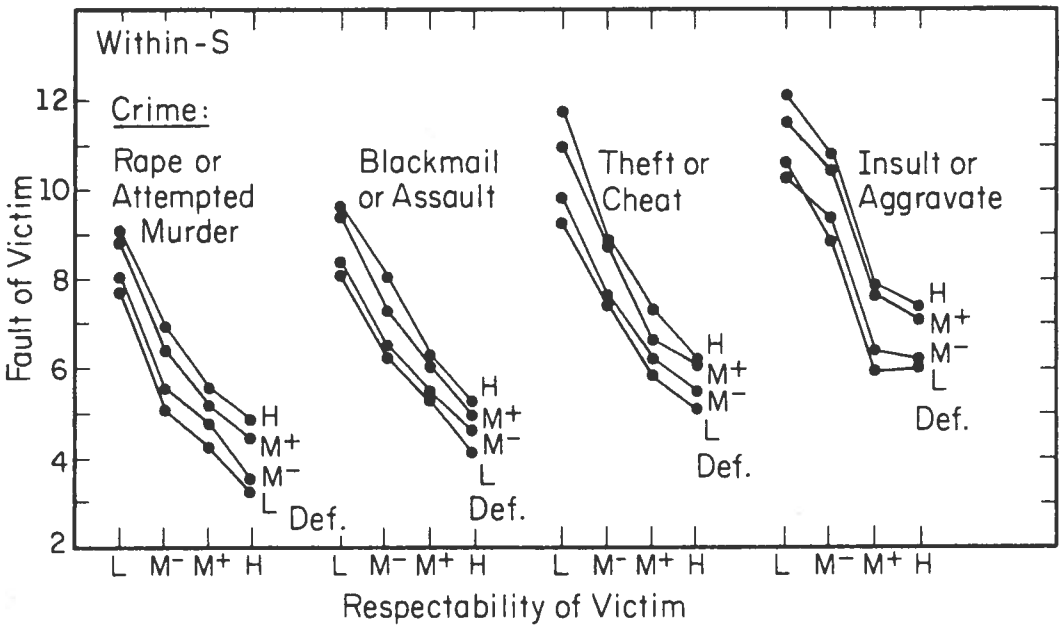


FIG. 17.24. Mean judgments of fault of victim as a function of victim's respectability in within-subject design, with a separate curve for each level of respectability of the defendant and a separate set of curves for each level of crime. From Birnbaum (1980c).

fore appears that the conclusion of Jones and Aronson (1973) can be reversed by changing from a between-subject to a within-subject design.

One can understand the finding that results change for between- vs. within-subject designs in this case by realizing that *in the between-subjects design, the stimulus and the context are completely confounded.* It is like the old stand-up joke:

Person 1. "How's your wife?"
Person 2. "Compared to what?"

Similarly, in a between-subjects experiment, the judge may ask him or herself, "How much at fault is this (former) virgin for her own rape? Compared to what? Compared to *other virgins,* perhaps."

Figure 17.25 shows a range-frequency analysis, assuming that virgins on the average are perceived to be more "innocent" than divorcées. In this analysis, a *raped virgin* is *more* innocent than a *raped divorcée* (see arrows on abscissa). But, she will be rated *less* innocent (more at fault) because *relative to the distribution of virgins,* a raped virgin is less innocent than a divorcée is *relative to the distribution of divorcées.* The curves show the predictions of range-frequency theory applied to the (presumed) distributions for this social judgment task. This interpretation explains how within- and between-subjects designs can give different results.

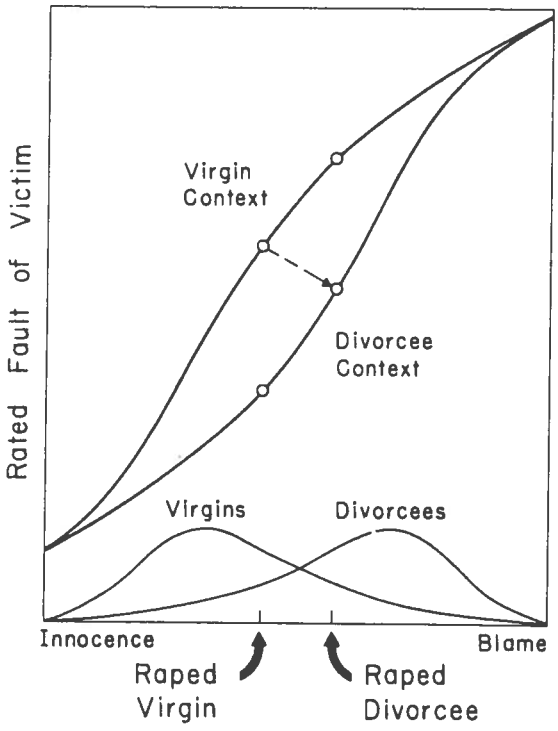


FIG. 17.25. A range-frequency analysis of social judgment in between-subject designs. Stimulus value and context are confounded in a between-subject design. In this case, the raped virgin is psychologically *less* at fault than a divorcée, but rated *more* at fault in a between-subjects design. Distributions show assumed contexts for different groups; curves are based on range-frequency theory. From Birnbaum (1980c).

It is not unusual in a between-subject design for the smaller stimulus to be judged greater. For example, in Figs. 17.2 and 17.3, the 27 dot pattern (in the positive context) is judged darker than is the 40 dot pattern (in the negative context). However, *within* each context (i.e., within subjects), the 40 dot stimulus is always judged darker. Therefore, comparing judgments of stimuli between groups of subjects who experience different contexts can be misleading.

Representative Design

Brunswik (1956) contended that the effects of psychological variables could depend on the experimental design—that is, on the range, spacing, and intercorrelations of the independent variables. Brunswik was concerned because in the natural environment, variables are correlated; whereas, in systematic research, variables are made independent in order to test theories of causation. He argued that if the correlation between variables is crucial to the subjects' performance and if the correlation is altered in the lab, then the results of systematic experiments would not generalize beyond the lab.

As an alternative to systematic design, Brunswik (1956) suggested representative design (see also Petrinovich, 1979, for a recent review). In representative design, the experimenter surveys the ecology (the subject's environment) and attempts to sample from it representatively. Brunswik also introduced the idea of *hybrid* design, discussing an example in which a factorial design of facial features (with schematic faces) was altered to produce a more representative correlation between two facial features.

Representative design is based on two ideas. First, it is based on the theory that the results of psychological investigations depend on the experimental design. This contention is reasonable in light of experimental evidence, including that presented in Section D. Second, it is based on the assumption that the foundation of generalization is representative sampling. In order to generalize from a political survey to an election, for example, pollsters attempt to obtain a representative sample. Statistical theory gives a rational basis for making inferences about population parameters on the basis of statistics computed from random samples. Problems of attempting to generalize from biased samples are well-known. Representative design emphasizes that treatments and situations should be sampled as well as subjects.

From these two ideas, Brunswik (1956) argued that psychologists should be willing to sacrifice experimental control for the sake of representativeness. He contended that systematic experiments are necessarily nonrepresentative because they are designed to unconfound variables that are actually confounded in nature.

There are two major problems with representative design. First, without experimental control, inferences of causation are unsound and dangerous. Second, the key element for generalization is not representative sampling but theory.

Let us consider, for example, a possible representative design for the study of health as a function of medical care. In each 6 month interval, the number of visits to a doctor and the state of the patient's health are recorded. In a representative design, people who see doctors more often have poorer health. This apparent harmful effect would persist even if the patient's diagnosis was partialed out; for example, among cancer patients, those seeing the doctor more often are in worse health. Only in a *very peculiar and nonrepresentative sample,* in which patients are randomly assigned to treatment vs. placebo conditions, is it possible to detect a beneficial effect of modern medicine. This example shows that correlations in representative samples can show relationships *opposite* of the direction of causation as inferred from systematic experimental research.

Reviewing the history of science, one finds many good examples of research that could not have succeeded with representative design. For example, the development of penicillin, television, the electric generator, Mendel's laws of genetics, and atomic and subatomic theory could not have occurred in representative research. These successes occurred because scientists were able to control simple situations in their laboratories and create new situations that do not occur in nature. It is difficult to find good counterexamples, where representative observation led to important results. The contributions of Jenner and Semmelweiss may fall in this category, but even their discoveries, which at first were based on observation of correlations, were doubted until verified by systematic research.

Two Notions of Generalization: Sampling vs. Theory

Let us consider two notions of generalization for an example experiment. An experimenter has tested a new drug using five fixed levels of concentration: .01, .02, .04, .08, and .16 moles/liter. Of the rats who have ingested lg per kg body weight of this mixture, the percentage who die within 1 hour is 2.5%, 16%, 50%, 84%, and 97.5% for the five levels of increasing dosage. According to the logic behind representative design, there is no basis for generalizing to levels intermediate in value or beyond the levels tested. There is also no basis for generalization from rats to humans. However, few people would be willing to ingest lg per kg body weight of a .32M concentration of this drug. The reason for this reluctance is that one generalizes to levels not tested from a *theory,* such as the following:

$$\text{Proportion killed} = F\ (\text{dosage}) \qquad \text{(E.1)}$$

where F is a monotonic function. We also theorize that if the drug kills rats, it would likely kill humans even though the drug has never been tested on humans before. This prediction is based on the very primitive theory that what kills one mammal is likely to kill another, a theory supported by considerable evidence. Furthermore, we might be willing to fit F to a particular function (cumulative

normal log dose) and predict that if a concentration of .0566 moles/liter were used, about 69% of the rats would die. Such prediction and extrapolation is based on *theory*, not representative sampling.

Let us return to the problem of psychophysical judgment and examine the consequences of different design strategies. Suppose we were interested in obtaining judgments of the heaviness of lifted weights. A standardized design would use a given stimulus range with a given stimulus spacing (a large range with geometric spacing has been advocated). A between-subjects design would present each subject with only one weight to lift. A representative design would survey the objects lifted in everyday life. This could be done by following people around and asking them to judge the heaviness of every object they happened to lift. The experimenter would then record the weights and other characteristics of the objects. By means of multiple regression, one would establish the effect of weight, size, etc. on heaviness judgments.

What can be made of the equation that predicts heaviness as a function of weight and other characteristics of the objects? Providing it is based on sufficient data, the function for weight obtained from this study can be used to give a reasonable statistical estimate of heaviness judgments as a function of weight in the population of objects from which the experiment can be considered a random sample. One *cannot* generalize beyond this population to other populations. Thus, representative design holds the context fixed to the context to which generalization is desired, and it provides no basis for generalization beyond the context studied. In order to generalize to all contexts, systematic manipulation of the contexts and development of contextual theory are required.

Systextual Design

Systextual design refers to systematic manipulation of the context (Birnbaum, 1975). The logic of systextual design is based on two premises: (1) it is necessary to manipulate the context in order to learn its effect; (2) one needs to develop a theory of context in order to generalize across contexts.

In the physics example, the approach of systextual design would be to develop a theory of the friction and thereby to predict observed departures from simple physical laws. By means of the theory, one could extrapolate to the frictionless situation or predict the results in a friction-filled one.

In the case of psychophysics, Parducci's research can be seen as an example of systematic manipulation of the range, stimulus spacing, frequency, response procedure, and so on. Parducci's range-frequency theory provides the possibility of predicting the judgment of a stimulus in any context—not just the standard context, or a subject's personal context, or the ecological context—but in principle it allows prediction to results across different contexts. Birnbaum (1974c) has shown how one can use range-frequency theory to derive a psychophysical scale from contextual effects in a fashion that demonstrates the invariance of the derived scale as a byproduct of fitting the theory.

In Birnbaum's (1974c) experiment, subjects judged the magnitude of numbers presented in one of nine stimulus spacings. Because the stimulus end points are fixed and because the psychophysical function was assumed to be strictly monotonic and error free, Birnbaum's (1974c) development of range-frequency theory yields:

$$\mathbf{C}_{ik} = aF_k(\Phi_i) + s_i \tag{E.2}$$

where $\mathbf{C}_{ik}$ is the rating of stimulus i in context k, $F_k(\Phi_i)$ is the cumulative proportion of stimuli less than Φ_i in context k, a is the weight of the frequency principle, and s_i is the scale value of stimulus i (calibrated on a linear scale). Once a has been estimated (which can be done using multiple linear regression with dummy variables for s), the equation can be written:

$$s_i = \mathbf{C}_{ik} - aF_k(\Phi_i) \tag{E.3}$$

Thus, by subtracting $aF_k(\Phi_i)$ from each category rating, it should be possible to derive a scale of psychological value that is the same for all contexts. Plotting $\mathbf{C}_{ik} - aF_k(\Phi_i)$ versus Φ_i should produce a set of curves for different contexts that all coincide, as in Birnbaum (1974c, Fig. 5).

Another example of systextual design is given in Birnbaum (1975). Judges were asked to press down on one end of a lever, lifting a weight at some distance from the fulcrum, and judge the force required to do so. The judge who understands the physics of the lever should expect the force required (to lift a weight) to vary directly with the distance of the weight from the fulcrum. However, in the systextual design, different weights were used to produce different correlations between force and position for different groups. A factorial design of force and position was embedded in an overall positive, negative, or zero correlation. In accord with Birnbaum and Veit (1973) and Birnbaum, Kobernick, and Veit (1974), it was theorized that judgments reflect a contrast between required force and expected force and that expected force depends on both the position and the subjective correlation between force and position. The model can be written

$$\mathbf{E}_{ijk} = Q_i - P_j R_k \tag{E.4}$$

where $\mathbf{E}_{ijk}$ is the judgment of the effort required, P_j is the position (distance from the fulcrum), R_k is the subjective correlation in context k, and Q_i is the effect due to actual force required. The results were consistent with the model and showed that the effect of position can indeed be reversed by reversing the correlation between force and position.

This experiment together with those of Birnbaum and Veit (1973) and Birnbaum et al. (1974) show that Brunswik was justified in his concern that the correlations among variables can affect the results of psychological experiments. Indeed, the effect of position can be reversed by changing its correlation with the variable to be judged. However, contrary to Brunswik's contention, it is possible to nest a systematic factorial design inside an overall correlation between var-

iables and to systematically manipulate the overall correlation. Furthermore, by means of such systextual design, it is possible to demonstrate the effect of the correlation and to develop a theory that permits generalization across correlations.

In summary, systematic design and representative design both hold the context fixed, and between-subject design confounds the context and the stimulus. Therefore, these designs do not permit tests of the empirical propositions upon which they are based. Systextual design calls for systematic manipulation of context and development of theory for generalization across contexts.

F. METHODOLOGY: ON MODEL TESTING AND MEASUREMENT

Conclusions regarding psychophysical processes are no better than the experimental, theoretical, and methodological foundations upon which they are based. In the study of psychophysical theories, it is useful to examine the logic of model testing and measurement carefully.

Anderson's functional measurement approach has had great impact on recent developments in psychophysical theory. The approach has many strong points in comparison with certain other approaches that have been well-expounded elsewhere (Anderson, 1970, 1977, 1979; Birnbaum, 1973, 1974b) and need not be repeated here. Instead, this section takes a critical look at the logic of functional measurement, from the skeptic's point of view. It is hoped that progress can be made by working to detect and strengthen weaknesses.

The following subsections review two substantive issues, impression formation and the size-weight illusion, to illustrate how weaknesses in the application of functional measurement led to the erroneous conclusion that these two processes could be represented by a parallel-averaging model. These two issues have been cited by Anderson (1979) to illustrate advantages of functional measurement, but they also serve well to illustrate limitations of the approach. Papers that have proposed methods to remedy defects in functional measurement are reviewed to show that previous conclusions regarding impression formation and the size-weight illusion do not stand up under improved experimental and analytical methods.

Six problems with simplistic applications of functional measurement are discussed. Several of these issues have been acknowledged by Anderson, but they have not been given sufficient attention. The following conclusions are discussed:

1. The fit of a model does not simultaneously validate the model, stimulus scale, and response scale.
2. Functional measurement is not a "neutral judge" between category rating and magnitude estimation.

3. Agreement of estimated scale values across tasks does not validate functional measurement.
4. Marginal means may not be linearly related to scale values even if the additive model fits the data (when the experiment lacks constraint).
5. The logic of two-stage integration analysis is inconsistent.
6. Methods involving the use of scale convergence and scale-free tests yield results that contradict previous conclusions from scale dependent research regarding the size-weight illusion and impression formation.

Parallelism Test

In functional measurement, a key method of analysis is the use of factorial design and analysis of variance. For example, the subject could be asked to judge the "average" sensation produced by two stimuli, A_i and B_j, where A_i and B_j have been factorially combined. The responses are plotted as a function of A_j with a separate curve for each level of B_i. Parallelism of the curves is equivalent to zero interaction between A and B.

In terms of the outline in Fig. 17.7, a set of premises that lead to parallelism can be listed as follows:

1. stimulus independence (e.g., s_{A_i} is independent of j)
2. $\Psi_{ij} = \dfrac{w_0 s_0 + w_A s_{A_i} + w_B s_{B_j}}{w_0 + w_A + w_B}$
3. $\mathbf{R}_{ij} = a\,\Psi_{ij} + b$

where s_{A_i} and s_{B_j} are scale values for the rows and columns, respectively, s_0 is the scale value of the initial impression and w_0 is its weight, w_A and w_B are the weights of the row and column factors, respectively, Ψ_{ij} is the subjective impression, $\mathbf{R}_{ij}$ is the overt response, and a and b are constants. Premise 2 is called the parallel-averaging model.

This model predicts that when $\mathbf{R}_{ij}$ is plotted against the column marginal mean ($\bar{\mathbf{R}}_{\cdot j}$) with a separate curve for each row, the curves should be linear and parallel. Thus, if the curves are *not* parallel, one should question the premises. If the curves *are* parallel, the premises can be retained.

However, parallelism does not validate the model, stimulus scale, and response scale all at once. True conclusions can be deduced from false premises. There are many sets of premises from which parallelism could be deduced. Some of these alternatives are shown in Table 17.2.

For example, the model could be multiplicative, and the judgment function could be logarithmic. It follows that $\mathbf{R}_{ij} = a\log \Psi_{ij} + b = a\log(s_{A_i} s_{B_j}) + b = a\log s_{A_i} + a\log s_{B_j} + b = s^*_{A_i} + s^*_{B_j} + b$ where $s^*_{A_i} = a\log s_{A_i}$. Therefore, a multi-

TABLE 17.2
A Few Ways to Explain Parallelism

Theory[a]	*Psychophysical Function* H	*Combination Function* C	*Judgment Function* J
1	(independence)	$\Psi_{ij} = s_{A_i} + s_{B_j}$	$\mathbf{R}_{ij} = a\Psi_{ij} + b$
2	(independence)	$\Psi_{ij} = s_{A_i} s_{B_j}$	$\mathbf{R}_{ij} = a\log\Psi_{ij} + b$
3	(independence)	$\Psi_{ij} = \sqrt{s_{A_i}^2 + s_{B_j}^2}$	$\mathbf{R}_{ij} = a\Psi_{ij}^2 + b$
4	$s'_{A_{(ij)}} = s_{A_i} + ks_{B_j}$ $s'_{B_{(ij)}} = s_{B_j} + ks_{A_i}$	$\Psi_{ij} = s'_{A_{(ij)}} + s'_{B_{(ij)}}$	$\mathbf{R}_{ij} = a\Psi_{ij} + b$

[a] Theory 1, 2, 3, and 4 can be titled the additive model, multiplicative model, Pythagorean model, and change of value model, respectively. These are some of the many alternative representations of parallelism.

plicative model could produce parallel data if the J function is logarithmic. Therefore, parallelism does not establish the validity of the response scale unless the additive (or parallel-averaging) model is assumed. Parallelism does not test the validity of the additive model unless the linear J function is assumed. Some additional constraint is needed (beyond arbitrary stipulation) to specify the functions of functional measurement.

Suppose the curves are nonparallel. How can nonparallelism be interpreted? There are two cases. In the first case, it may be possible to reject the additive or parallel-averaging model on the basis of the ordinal information in the data, when the data systematically violate independence or double cancellation (Krantz & Tversky, 1971). In the second case, the numerical data are not parallel, but they can be rescaled to parallelism by means of a monotonic transformation. In this case, it is not possible without additional constraint to specify whether the nonparallelism is due to a nonadditive integration function or to a nonlinear J function. This point is expanded upon in the discussion of impression formation and the size-weight illusion.

A debate between proponents of conjoint measurement (Krantz et al., 1971) and Anderson occurred over the propriety of rescaling data from Sidowski and Anderson (1967) who found an interaction between cities and occupations for judgments of job desirability. Krantz et al. (1971) rescaled the mean judgments to parallelism and argued that the interaction analyzed by Sidowski and Anderson could be without psychological significance.

Thus, if the data are parallel, many combinations of $J[C(s_i, s_j)]$ are possible. If the data are not parallel, but can be rescaled to parallelism, many combinations of $J[C(s_i, s_j)]$ are still possible. Some have concluded that the parallelism test is therefore nondiagnostic. However, it does have value because the realm of possibilities for J and C in the parallel and nonparallel cases are different.

Is Functional Measurement a Neutral Judge?

Anderson (1972, and this volume) argued that functional measurement serves as a "neutral judge" between magnitude estimation and category rating. This contention was illustrated with reference to an experiment by Weiss (1972), who obtained magnitude estimations and graphic ratings of the "average" darkness of two gray chips. Anderson argued that because subjects were instructed to "average," one should postulate a parallel-averaging model. The rating data were approximately parallel whereas the magnitude-estimation data showed bilinear divergence. Anderson concluded that ratings are "valid," but magnitude estimations are "biased and invalid."

At least three other studies have directly compared magnitude estimation with rating methods in situations employing factorial designs and the same task. Sarris and Heineken (1976) used these two procedures for the judgment of heaviness of size-weight blocks. Curtis and Rule (1978) extended the study of Weiss (1972) to include "average" lightness and darkness using the two-response procedures. Veit (1978) obtained ratings and magnitude estimations of "differences." Marks (1979) had subjects rate the "overall loudness" of a multicomponent tone using magnitude estimation and a graphic rating procedure. In each case, the effect of the response procedure was represented by changes in the J function. In each case, magnitude estimations were positively accelerated relative to ratings. In three of the studies, the *assumption* of an additive (or subtractive) model would lead to the conclusion that the J function for ratings is nearly linear, and the J function for magnitude estimation is positively accelerated. Marks (this volume) found that to *assume* the additive model for loudness summation required negatively accelerated J functions for both response procedures.

The size-weight experiment of Sarris and Heineken (1976) obtained results similar to those of Weiss (1972). Using magnitude estimations, the data were nearly consistent with a geometric averaging model (multiplicative). If one *assumes* that grayness "averaging" and the size-weight illusion can be represented by the parallel-averaging model (as did Anderson, 1972), one would conclude that the J function for magnitude estimation is nonlinear. If one were to *assume* that grayness "averaging" and the size-weight illusion should be represented by a geometric averaging model, however, one would conclude that magnitude estimations are "valid" and that ratings are "biased." Thus, the situation is circular. In order to decide on the "valid" scale, one must *assume* the model. To choose the appropriate model, one must *assume* the "valid" scale. Birnbaum and Veit (1974b) termed this problem "scale-dependence," in which the conclusions regarding the model depend on the arbitrary decision to place faith in the particular dependent variable and the particular context that led to either parallelism or bilinearity.

To argue that Weiss (1972) has shown magnitude estimation to be biased requires the assumption either that ratings are valid (making the argument com-

pletely circular) or that the appropriate model is additive (which is semicircular). Were one to assume that the model is multiplicative (as in a geometric averaging model), it would be concluded that magnitude estimation is "valid" and category rating is "biased and invalid." Unless the model is assumed, the conclusion is scale dependent (Birnbaum & Veit, 1974b); unless the scale is assumed, the conclusion is model dependent.

Therefore, if functional measurement was truly neutral (i.e., did not prejudge the validity of the response scale or model), then the experiments cited by Anderson would be inconclusive on the question of the "validity" of ratings and magnitude estimations in experiments like that of Weiss (1972) and Sarris and Heineken (1976).

Anderson (1977) argued that the circularity of his conclusion regarding magnitude estimation can be ameliorated by considering the success of the parallelism test using category ratings in impression-formation research. However, it is shown next that the early work in impression formation was inadequate and reached erroneous conclusions.

Impression Formation

Perhaps no paper has been as often cited to illustrate Anderson's approach as his first article on impression formation (Anderson, 1962). Anderson (1962) had 12 subjects judge the likeableness of hypothetical persons described by sets of adjectives, using a 20-point rating scale. Anderson's theory of these data can be written as follows:

$$\Psi = \sum_{i=0}^{k} w_i s_i \Big/ \sum_{i=0}^{k} w_i, \tag{F.1}$$

where Ψ is the integrated impression, w_i and s_i are weight and scale value of adjective i, and w_0 and s_0 are the weight and scale value of a postulated initial impression.

The adjective combinations were generated from a factorial design. Anderson noted that: (1) if the weights are independent of scale value; (2) if the scale value of each adjective is independent of the other adjectives with which it is paired; (3) if the response scale is "valid" (i.e., J is linear); and (4) if impressions are governed by Eq. F.1, then the data would show parallelism, and there would be nonsignificant interactions among the adjective factors in analysis of variance. Anderson (1962) found that the majority of his 12 subjects had nonsignificant interactions.

What can be concluded from the experiment? It has been contended that the fit of the model simultaneously "validates" the stimulus scale, response scale, and model all at once. However, as just noted, this view is oversimplified, for the

conclusion in principle is scale dependent. Furthermore, it can also be shown that the basic finding does not replicate.

A Divergent Finding. Anderson's (1962) conclusions regarding impression formation were challenged by Birnbaum (1974a, Exp. 1), who obtained a large divergent interaction for ratings of likeableness. The left of Fig. 17.26 shows mean judgments of likeableness as a function of one adjective with a separate curve for each level of the other. The means in Fig. 17.26 are averaged over six different sets of adjectives. Each off-diagonal point is the average of 600 judgments by 300 subjects. Results for individual adjectives are given in Birnbaum (1974a, Fig. 2). Although there are other aspects of individual adjective and subject data that are of interest (see Birnbaum, 1974a), the divergence shown in Fig. 17.26 was characteristic of individual data.

The interaction obtained by Birnbaum (1974a, Exp. 1) reopens all of the issues of impression formation. Nonparallelism indicates that impression formation may violate the parallel-averaging model, that J could be nonlinear, that the scale values of the adjectives could change as a function of the adjectives with which they are paired, or any of a number of other possibilities.

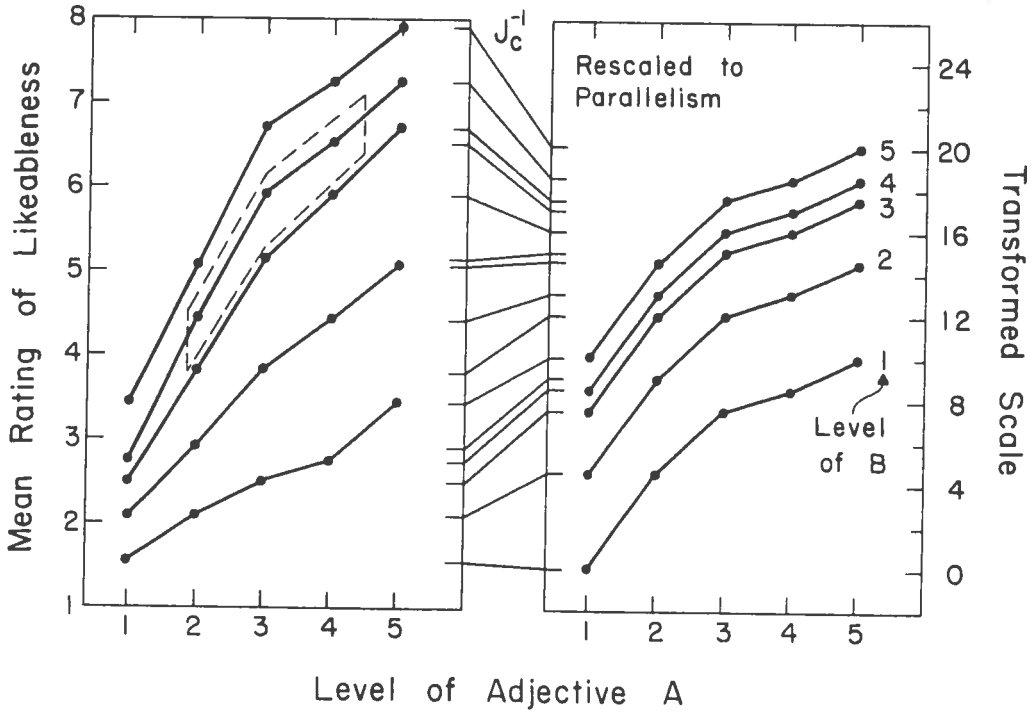


FIG. 17.26. Mean ratings of (combined) likeableness as a function of the level of likeableness of adjective A, with a separate curve for each level of adjective B. Mean ratings (left panel) are not parallel, but they can be monotonically rescaled to parallelism, as shown in panel on the right. Dashed box shows that domain of Lampel and Anderson (1968) was small in comparison with that of Birnbaum (1974a). From Birnbaum (1980b).

A more prosaic possibility is that the difference in results between Anderson (1962) and Birnbaum's Experiment 1 (1974a) is due to differences in experimental procedure. It was once argued that the parallel-averaging model held, but only under very special experimental conditions. Birnbaum (1974a) systematically varied the experimental procedures to see if any of the following manipulations would remove the interaction experimentally: The category scale labels were reversed, a 20-point anchored scale was used, a line mark response scale was used, a matching procedure (analogous to method of adjustment) was used to eliminate numerical response, and "equal accuracy and importance" instructions were tried. No evidence was found that the interaction could be removed by these variations in the experimental procedure. The divergent interaction appeared with all of these procedures.

Rescaling to Parallelism. In order to rescue the parallelism-predicting averaging model (Eq. F.1), it is possible to rescale the data in the left of Fig. 17.26 to parallelism, as shown on the right of Fig. 17.26. The scale values 0, 4.5, 7.5, 8.5, and 10 for the five levels can be added or averaged to produce the parallel curves on the right of Fig. 17.26. The rank order of the means can be perfectly reproduced by the parallel-averaging (or additive) model. Therefore, the divergence in the left of Fig. 17.26 could be explained either by the assumption that C (the model representing impression formation) contains an interaction, or that J (the judgment function) is nonlinear and C is a parallel-averaging (or additive) model.

The Rescaling Debate

It is instructive to discuss a possible debate between a mentalist measurer and a behaviorist model tester concerning the data in Fig. 17.26. The behaviorist declared that the data on the left of Fig. 17.26 allow one to refute Anderson's (1962) additive (or parallel-averaging) model of impression formation. The mentalist declared the data ordinally consistent with the parallelism model and used the model to measure scale values for the adjectives, as on the right of Fig. 17.26.

"But you're *assuming* the model I just disproved!" the model tester exclaimed.

The mentalist replied, "The data can be rescaled to additivity, so I see no problem."

"But the violation of parallelism in the raw data is inconsistent with the additive model," the behaviorist noted.

"Only if you assume that J is linear," the measurement mentalist said, becoming edgy. "It is simpler and more reasonable to assume that the parallel-averaging model described the combination process. After all, subjects can't be trusted to do any more than rank order their impressions."

The behaviorist grew confident, "I see no reason to postulate a J function at all. I have operationally defined likeableness in terms of my rating procedure. As a behaviorist, I want a model that describes the raw data I obtained. The raw

data do not fit the additive model. As a model tester, I therefore reject the additive model in favor of a model with an interaction. You are fudging the data with your transformation. You are rescaling the data to fit the model and then trying to tell me that the model fits!''

The mentalist grew irritated and looked away as he said, ''I find no reason to reject the model if the data satisfy the ordinal (rank order) requirements. The data are perfectly additive, in the ordinal sense that $R_{ij} > R_{kl}$ whenever $s_i + s_j > s_k + s_l$. You assume that J is linear and are trying to reject the model for no valid reason. Category ratings are nonlinearly related to magnitude estimations. How can we be convinced that the ratings shouldn't be monotonically transformed?''

''Anderson validated ratings by showing that impression-formation data are parallel,'' the behaviorist replied weakly.

The mentalist looked him in the eye and snapped, ''Now you've contradicted yourself! You are *assuming* the model in order to *validate* the response scale you must assume for your test, and then you reject the model you assumed in the first place! Parducci has shown repeatedly that category ratings in two situations can be nonlinearly related if the stimulus distribution is changed. The change in stimulus distribution doen't change the rank order of the points, but it *does* change the apparent parallelism. Look at Fig. 17.20 in this chapter! You can't be sure that ratings are linearly related to subjective value because ratings in one context are nonlinearly related to ratings in another context, as in Fig. 17.21. Therefore, we must allow for nonlinear transformation of the data.''

Just then, an aged philosopher stepped up with a look of condescension and said: ''You two are arguing over a meaningless distinction. What difference does it make whether the interaction comes from C or J? You'll never be able to settle your dispute on empirical grounds, because the two theories are equivalent.''

The next subsections show that, contrary to the philosopher, it is possible to design new experiments that can test between the two theories, if one is willing to accept the principle of scale convergence and the logic behind the scale-free test.

Scale-Convergence Criteria

To decide whether the interaction shown in Fig. 17.26 was ''real'' (i.e., due to C in Fig. 17.7) or reflecting only response ''bias,'' (i.e., due to J) Birnbaum (1974a) advanced the criteria of stimulus and response scale convergence. The stimulus scale convergence criterion assumes that the likeableness scale values of adjectives should be independent of the task, which in this case was to judge ''differences'' in likeableness or ''combinations'' (integrated impressions). The response scale convergence criterion states that consistent principles determine the J function. It was postulated that if the same subjects used the same response procedure to judge the same stimuli on the same dimension presented in the same distribution, the J functions should be the same for both tasks (see Birnbaum, 1974a).

The stimulus scale convergence theory can be written as follows:

$$\mathbf{D}_{ij} = J_{\mathbf{D}}[s_j^* - s_i^*] \tag{F.2}$$

$$\mathbf{C}_{ij} = J_{\mathbf{C}}[\Psi_{ij}] \tag{F.3}$$

$$\Psi_{ij} = \frac{w_0 s_0 + w_A s_i + w_B s_j}{w_0 + w_A + w_B} \tag{F.4}$$

$$s_i^* = s_i \text{ (scale convergence)} \tag{F.5}$$

where $\mathbf{D}_{ij}$ and $\mathbf{C}_{ij}$ are the ratings of "differences" in likeableness between two different people, each described by an adjective, and overall (combined) likeableness of a person described by both adjectives. Equation F.5 explicitly assumes that $s^* = s$, i.e., that the scale value for the likeableness of an adjective is the same for both "combination" and "difference" tasks. The response scale convergence theory would allow s^*_i and s_i to be different, but would assume that $J_{\mathbf{C}}$ and $J_{\mathbf{D}}$ have the same functional form, within a linear transformation.

The data of Birnbaum (1974a, Exp. 3) required rejection of both scale convergence criteria if the parallel-averaging model was assumed. However, both criteria could be retained if the parallel-averaging model was rejected. Figure 17.27 shows the results for the "differences" experiment (Birnbaum, 1974a, Exp. 3). Note that the data for "combinations" show a divergent interaction (left of Fig. 17.26) whereas the data for "differences" are nearly parallel. To retain the premise that $J_{\mathbf{D}}$ and $J_{\mathbf{C}}$ are both linear would require the rejection of the parallel-averaging model. To retain the subtractive model of "differences"

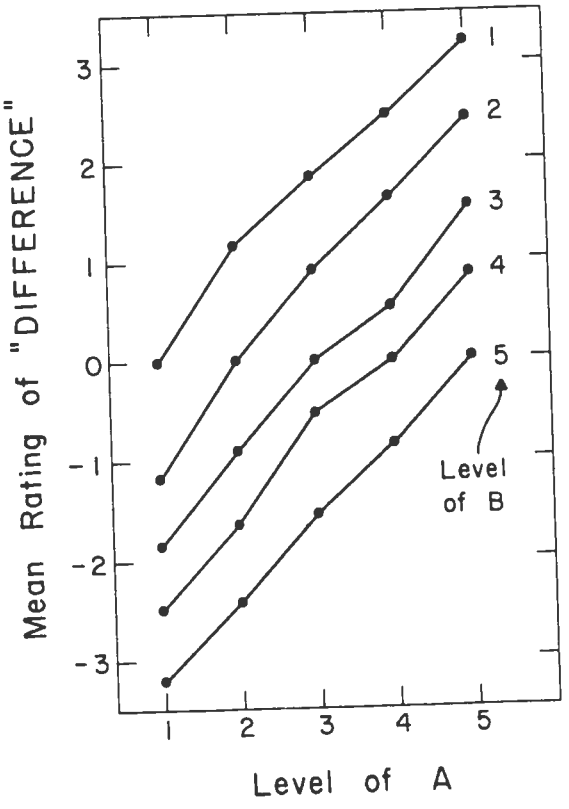


FIG. 17.27. Mean rating of "difference" in likeableness between two persons, each described by one adjective. From Birnbaum (1980b).

and the parallel-averaging model of "combinations" would require the rejection of the premise that J_D and J_C were of the same form. Assuming both models, J_C would be positively accelerating, and J_D would be linear.

However, if both models are assumed, the data violate stimulus scale convergence. To understand this, it is instructive to derive the ordinal constraints imposed by both the data and the additive model on the scale values and illustrate the systematic violations. Let $a = s_2 - s_1$, $b = s_3 - s_2$, $c = s_4 - s_3$, and $d = s_5 - s_4$ for the "combination" task. Let a^* to d^* be defined analogously for the "difference" task. The theoretical representations for the subtractive and additive models are shown in parentheses in Tables 17.3 and 17.4, together with the mean judgments.

By comparing the rank order of the "differences" with the theoretical representations in Table 17.4, it can be found that the scale for "differences" must satisfy the following: $0 < c^* < d^* < b^* < a^* < c^* + d^* < b^* + c^* < a^* + b^* < b^* + c^* + d^* < a^* + b^* + c^* < a^* + b^* + c^* + d^*$, which can be simplified, as follows:

$$0 < c^* < d^* < b^* < a^* < c^* + d^* \qquad \text{(F.6)}$$

TABLE 17.3
Mean Ratings of Likeableness[a]

	Level of A				
Level of B	*1*	*2*	*3*	*4*	*5*
1	1.54 (0)	2.10 (a)	2.50 ($a + b$)	2.76 ($a + b + c$)	3.45 ($a + b + c + d$)
2	2.10	2.92 ($2a$)	3.82 ($2a + b$)	4.44 ($2a + b + c$)	5.08 ($2a + b + c + d$)
3	2.50	3.82	5.15 ($2a + 2b$)	5.90 ($2a + 2b + c$)	6.72 ($2a + 2b + c + d$)
4	2.76	4.44	5.90	6 53 ($2a + 2b + 2c$)	7.25 ($2a + 2b + 2c + d$)
5	3.45	5.08	6.72	7.25	7.90 ($2a + 2b + 2c + 2d$)

[a] Each entry is the mean judgment of likeableness of a person described by both A and B. Each off-diagonal cell is averaged over six pairs of adjectives; 600 judgments from 300 subjects. Data from Birnbaum (1974a, Exp. 1). Algebraic symbols give additive representation, $C_{ij} = J_C[s_i + s_j]$, with $s_1 = 0$, $a = s_2 - s_1$, $b = s_3 - s_2$, $c = s_4 - s_3$, $d = s_5 - s_4$. Arrows represent inequalities showing that $a > b + c$ and $a > c + d$. Parallel-averaging model is equivalent to additive model in this case.

TABLE 17.4
Mean Ratings of "Differences"[a]

	Level of A				
Level of B	1	2	3	4	5
1	0	1.18 (a^*)	1.86 ($a^* + b^*$)	2.49 ($a^* + b^* + c^*$)	3.20 ($a^* + b^* + c^* + d^*$)
2	−1.18	0	.92 (b^*)	1.64 ($b^* + c^*$)	2.43 ($b^* + c^* + d^*$)
3	−1.86	−.92	0	.53 (c^*)	1.54 ($c^* + d^*$)
4	−2.49	−1.64	−.53	0	.85 (d^*)
5	−3.20	−2.43	−1.54	−.85	0

[a] Each number is the mean judgment of "difference" in likeableness, A-B. Each cell is averaged over six pairs of adjectives; 180 judgments from 90 subjects. Data from Birnbaum (1974, Exp. 3). Algebraic symbols give subtractive representation, $\mathbf{D}_{ij} = J_{\mathbf{D}}[s_j^* - s_i^*]$, with $s_1^* = 0$, $a^* = s_2^* - s_1^*$, $b^* = s_3^* - s_2^*$, $c^* = s_4^* - s_3^*$, $d^* = s_5^* - s_4^*$. Arrows represent inequalities showing that $a^* < b^* + c^*$ and $a^* < c^* + d^*$.

Notice that each difference between successive scale values is less than any two-step difference. A set of scale values satisfying Expression F.6 would be 0, 10, 18, 24, 31, where $a^* = 10$, $b^* = 8$, $c^* = 6$, and $d^* = 7$.

However, if the additive (or parallel-averaging) model is assumed, the rank order in Table 17.3 implies the following:

$$0 < c < d < b < b + c < a < b + c + d \tag{F.7}$$

Note that $a > c + d$ and $a > b + c$, but $a^* < c^* + d^*$ and $a^* < b^* + c^*$. These contradictions in ordinal relationships for differences in scale value imply that s is nonlinearly related to s^*. In particular, Expressions F.6 and F.7 imply that the scale values for the additive model of "combinations" are concave downwards relative to the scale values estimated from the subtractive model applied to "differences." Thus, even allowing $J_{\mathbf{C}}$ and $J_{\mathbf{D}}$ to be any monotone functions, it is not possible to retain the theory expressed in Eqs. F.2 through F.5.

In summary, to conclude that the parallel-averaging model underlies impression formation and the subtractive model represents "difference" judgment would require rejection of both stimulus and response scale convergence criteria in favor of the conclusion that there are two different scales of likeableness, s^* and s, where s^* (for "differences") is positively accelerated relative to s, and two different output functions, $J_{\mathbf{C}}$ and $J_{\mathbf{D}}$ for category ratings, where $J_{\mathbf{C}}$ is

positively accelerated, and J_D is nearly linear. The next subsection shows that it is possible to retain the scale convergence criterion and the subtractive model for stimulus comparison by rejecting the parallel-averaging model in favor of a configural-weight model.

Configural-Weight Model of Impression Formation

A simple configural-weight model can describe the data in the left of Fig. 17.26, using the scale values obtained from the fit of the subtractive model to "difference" judgments. The configural-weight averaging model assumes that the worst trait of a person receives extra weight. Figure 17.28 shows predictions for the configural-weight model using scale values of 0, 10, 18, 24, and 31 for the five levels of likeableness of the adjectives, assuming $s_0 = 17$. All weights were set to 1.0, but the lowest scale value in each set (which could be s_0 on some trials) was assumed to have a weight of 2.0. For example, the predicted value for cell (1, 5) would be $(2 \cdot 0 + 1 \cdot 17 + 1 \cdot 31)/(2 + 1 + 1) = 12$. The predicted value for cell (3, 3) would be $(2 \cdot 17 + 1 \cdot 18 + 1 \cdot 18)/(4) = 17.5$.

The predictions of the configural-weight model in Fig. 17.28 have the same rank order as the mean ratings in the left of Fig. 17.26, they show a similar pattern of divergent interaction, and they are based on the same scale values as for the subtractive model applied to "difference" judgments. The configural-weight theory differs from the parallel-averaging theory in that it postulates a

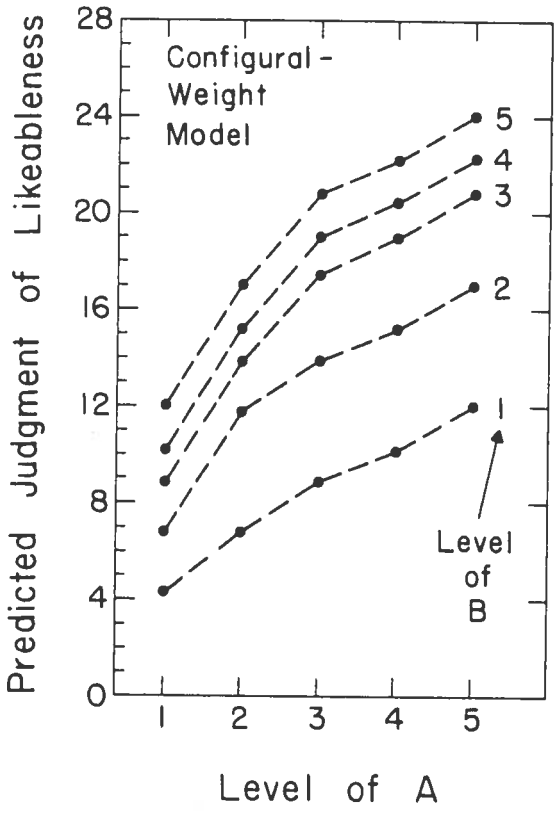


FIG. 17.28. Predicted likeableness of combinations, based on a configural-weight model using scale values from "difference" task (Fig. 17.27). Note that rank order and shape of curves is the same as in the left of Fig. 17.26. From Birnbaum (1980b).

'real'' psychological interaction between the adjectives: If a person has one bad rait, other traits will have less effect. (Configural-weight models are discussed in greater detail by Birnbaum [1974a] and Birnbaum & Stegner [1979], who also liscuss a competing differential-weight averaging model, which can also predict livergence).

Thus, the following theory remains consistent with the data:

$$\mathbf{D}_{ij} = J_{\mathbf{D}}[s_j^* - s_i^*] \tag{F.8}$$

$$\mathbf{C}_{ij} = J_{\mathbf{C}}[\Psi_{ij}] \tag{F.9}$$

$$\Psi_{ij} = I[s_0, s_i, s_j] \tag{F.10}$$

$$s_i^* = s_i \tag{F.11}$$

where I is an integration rule for impression formation that contains a ''real'' divergent interaction. Figure 17.28 shows that a configural-weight model can reproduce the rank order of the combination data using a single set of scale values for both ''differences'' and ''combinations'' and using functions for both $J_{\mathbf{C}}$ and $J_{\mathbf{D}}$ that are nearly linear. In conclusion, by rejecting the parallelism-predicting models of impression formation, it is possible to retain scale convergence. To retain scale convergence requires rejection of the parallelism-predicting models of impression formation.

Scale-Free Tests of Impression Formation

It could be argued that the stimulus and response scale convergence criteria should be rejected, rather than the parallel-averaging model. Thus, it could be argued that there are different J functions *and* different scale values for the ''difference'' and ''combination'' tasks. Of course, such an argument is complicated, for it provides no theory to explain the change in scale values or J function beyond perhaps a vague remark that judgment proceeds in stages, so ''why not insert a few more stages with internal transformations?''

In response to this possibility, Birnbaum, (1974a, Exp. 4) introduced the scale-free test. In the scale-free test of impression formation, subjects judge ''differences'' in likeableness between pairs of hypothetical persons, each described by two adjectives. For example, how much more would you like a person described as *loyal* and *understanding* than one described as *loyal* and *malicious?*

These judgments of differences between combinations, **DC**, can be represented by the model:

$$\mathbf{DC}_{ijkl} = J[\Psi_{ij} - \Psi_{kl}] \tag{F.12}$$

where J is a monotone function and Ψ_{ij} and Ψ_{kl} are the integrated impressions of likeableness.

If the parallelism-predicting model of impression formation is correct, then the following two judgments of ''differences'' should be equal:

1. *Loyal* (*L*) and *understanding* (*U*) vs. *loyal* (*L*) and *obnoxious* (*O*).
2. *Malicious* (*M*) and *understanding* (*U*) vs. *malicious* (*M*) and *obnoxious* (*O*).

According to Eq. F.12, the first "difference" can be represented $J[\Psi_{LU} - \Psi_{LO}]$, which according to the additive model can be written $J[s_L + s_U - s_L - s_O] = J[s_U - s_O]$. The second "difference" can be expressed $J[\Psi_{MU} - \Psi_{MO}] = J[s_U - s_O]$. Therefore, the judged "difference" between *loyal and understanding* compared with *loyal and obnoxious* should be *equal* to the judged "difference" between *malicious and understanding* compared with *malicious and obnoxious*. The alternative hypothesis, that the divergent interaction in the left of Fig. 17.26 is "real," predicts that the first "difference" is greater than the second "difference." This test between a null hypothesis of equality and a directional inequality assumes only that J is a strictly monotonic function. the test is termed "scale-free" because the conclusion regarding the additive model of impression formation is invariant with respect to strictly monotonic transformations of the "difference" judgments. All that matters is the rank order of the "difference" judgments.

The left side of Fig. 17.29 shows the result for a part of the scale-free test of

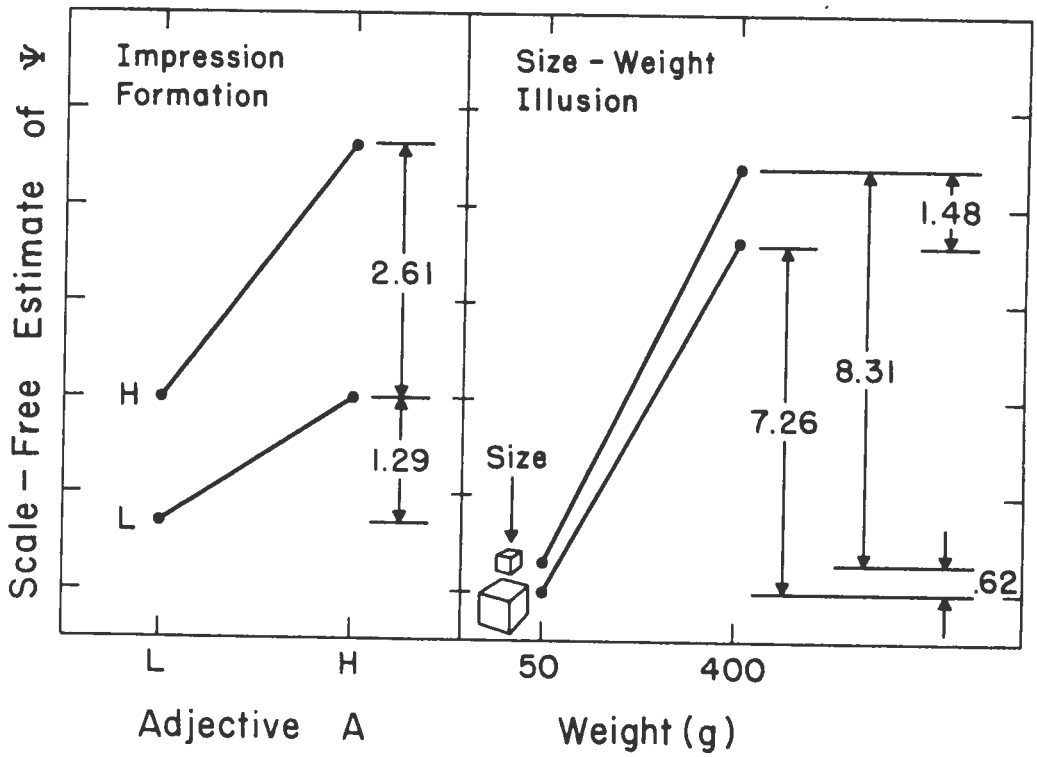


FIG. 17.29. Scale-free tests of additive (or parallel-averaging) model of impression formation and the size-weight illusion. No strictly monotonic transformation could rescale these data to parallelism. Subjects judge the "difference" between HH and HL to be 2.61, but they judge the "difference" between LH and LL to be only 1.29. Additive model requires that these two judgments be equal. From Birnbaum (1980b).

Experiment 4 of Birnbaum (1974a) for impression formation. For simplicity, only a portion of the design is presented here, and the means have been averaged over subjects and adjective replicates. (A constant has been added so that a zero "difference" is 0.) Figure 17.29 shows the divergence characteristic of ratings of likeableness in Fig. 17.26. Thus, the rank order of "difference" judgments *in this case* is predictable from differences in mean rating. It should be clear that the divergence in Fig. 17.29 would persist under any strictly monotonic transformation of the "difference" judgments.

The assumption of Eq. F.12 can be justified on the basis of Sections B and C in this chapter. However, even if Eq. F.12 were replaced with a ratio model of comparison, the conclusions regarding the additive model of impression formation would be the same (Birnbaum, 1979). Transformation to a ratio model of comparison would require exponential transformation, which would increase the divergence.

In sum, the scale-free test refutes the parallel-averaging model of impression formation.

Size-Weight Illusion

Anderson's (1972) experiment on the size-weight illusion has also been used frequently (e.g., Anderson, 1977, 1979) to illustrate his views on functional measurement and data analysis.

Anderson's (1972) data are shown in the left side of Fig. 17.30. Judgments of

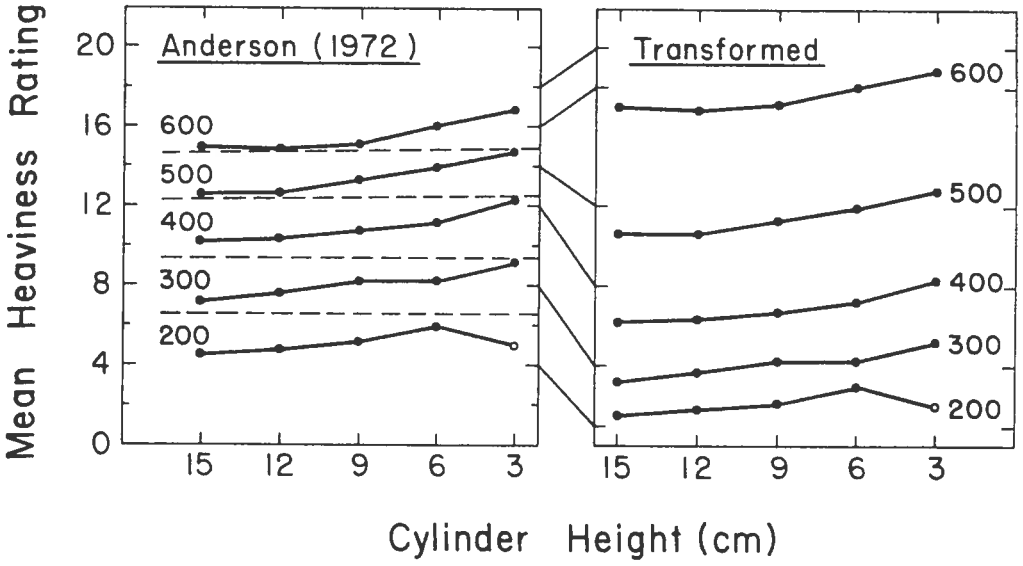


FIG. 17.30. Left panel shows mean judgments of heaviness from Anderson's (1972) study of the size-weight illusion. Dashed lines between curves show that the rank order of means does *not* constrain the scale values of heaviness. Entire curves can be shifted up or down without changing rank order of data. Points on right are monotonically related to original, and they fit the additive model equally well, yet transformed data imply scale values for heaviness that are a positively accelerated function of weight. From Birnbaum (1980b).

heaviness are plotted as a function of the size of the cylinder, with a separate curve for each level of weight. Although the interaction was statistically significant, Anderson attributed it to an experimental problem (he discarded the open point) and assumed the data were essentially parallel. If the model is additive, he reasoned, parallelism "validates" the response scale. Furthermore, he assumed that the spaces between the curves provided a scale of heaviness. Inasmuch as the spaces between the curves (representing 100 g increments) decreased as weight increased, Anderson concluded that heaviness is a negatively accelerated function of weight. Because magnitude-estimation experiments yielded exponents for heaviness greater than one, Anderson concluded that magnitude estimation must be "biased and invalid."

However, the data shown in Fig. 17.30 do not warrant such strong conclusions. Even if it were granted that the size-weight illusion is additive (which is disputed later), the data in Fig. 17.30 do not determine scale values for heaviness. To see why this is so, study the dashed lines in the left of the figure. These lines show that the data can be monotonically rescaled to many other equally additive solutions by shifting entire curves up or down. In other words, the rank order in Fig. 17.30 places virtually no constraints on the scale values for weight. For example, the right panel of the figure shows that the data can be rescaled to yield a positively accelerated psychophysical function for heaviness, in which the distances between successive curves actually *increase* with increasing weight.

Unconstrained Scale Values

Although the additive model constrains scale values to interval scale uniqueness *in principle,* and although parallelism demonstrates linearity of J *in principle* (if the additive model is assumed), the experimental design of Anderson (1972) fails to provide enough constraint either to test the linearity of J or to constrain scale values for heaviness. In this case, the failure to ask if the data allowed one to *refute* the possibility of a positively accelerated heaviness scale led to the unfounded conclusion that the additive model for Anderson's (1972) size-weight data was inconsistent with the heaviness scales from magnitude estimation.

If one is willing to consider multiplicative models, then even had the data been highly constrained, perfectly parallel, and shown a log-function for heaviness, the experiment could not *in principle* yield the conclusion that magnitude estimation is "biased and invalid." Suppose both the size-weight illusion and "averaging" task data were perfectly parallel and suppose the scale values were identical. Exponential transformation on both sets of data would yield perfect bilinearity, which would be deemed consistent with a multiplicative model. Furthermore, the so-called "cross-task validation" would still work. As Birnbaum and Veit (1974a, 1974b) predicted, and as Sarris and Heineken (1976) observed, when ratings fit additive (or subtractive) models, magnitude estimations tend to fit multiplying (or ratio) models. Anderson (1972) recognized the possibility of

data transformation to other models but failed to point out that if magnitude estimations are exponentially related to ratings, then his conclusion regarding the validity of magnitude estimation would be an arbitrary decision rather than an empirical finding.

Scale-Free Test of Size-Weight Illusion

Birnbaum and Veit (1974b) noted that the situation with respect to the size-weight illusion is perfectly circular in scale-dependent research, such as that of Anderson (1972). If the additive model were assumed and ratings were additive, then ratings would be "validated." If the ratio model were assumed and magnitude estimations fit this model (as in Sarris & Heineken, 1976), then magnitude estimation would be "validated." Unless one model or one response scale is assumed, no conclusion can be drawn.

In order to go beyond the circular situation of scale-dependent research, Birnbaum and Veit (1974b) applied the scale-free test of Birnbaum (1974a, Exp. 4) to the size-weight illusion. Birnbaum and Veit (1974b) asked subjects to judge the "difference" in heaviness between pairs of size-weight blocks. It was assumed that "difference" ratings can be represented as follows:

$$\mathbf{D}_{ijkl} = J[\Psi_{ij} - \Psi_{kl}] \tag{F.13}$$

where Ψ_{ij} is the heaviness of weight i in block j. It follows from an additive (or parallel-averaging) model that the "difference" in heaviness between two blocks of the same weight but different sizes should be independent of that weight (t_i). If $\Psi_{ij} = t_i + s_j$, where t_i is the weight and s_j is the size, then $\mathrm{D}_{ijil} = J[t_i + s_j - t_i - s_l] = J[s_j - s_l]$. Thus, the additive model implies that the magnitude of the illusion should be independent of weight. Similarly, $\mathrm{D}_{ijkj} = J[t_i + s_j - t_k - s_j] = J[t_i - t_k]$.

On the other hand, if the divergent interaction observed by Sjöberg (1969) is real, then the magnitude of the illusion should increase as a function of weight. Similarly, if the interaction is "real," then the difference in heaviness between two different weights in blocks of the same size should depend on the common size (see Birnbaum, 1974a, Fig. 5).

Because these predictions of equality or inequality hold for any strictly monotonic function J, the test is a scale-free test. In other words, the conclusion regarding the additive model for the size-weight illusion would be invariant with respect to strictly monotonic transformation of the data.

Birnbaum and Veit (1974b) found a systematic violation: The difference in heaviness between two blocks of different sizes was larger when both blocks weighed 400 g than when both blocks weighed 50 g. Similarly, the judged difference in heaviness between two blocks of the same size that weighed 50 and 400 g was ranked larger when the block was small than when it was large. Fifteen of the 16 subjects in Experiment 2 showed these trends in the crucial rank-order

test. The scale-free estimates of heaviness, derived from the subtractive model of comparison, are shown in the right of Fig. 17.29,where the numbers in the figure show the mean "difference" judgments.

Birnbaum and Veit (1974b) used a more extensive design than that just described. There were actually 21 size-weight blocks composed of seven levels of weight in blocks of three different sizes. These were factorially combined with four size-weight combinations (in the other hand) for comparison. The design was counterbalanced across the two hands, and there were two replicates, yielding a $2 \times 2 \times 4 \times (3 \times 7) = 336$-cell design. The 336 data points for each subject were rescaled to the following model:

$$J^{-1}(\mathbf{D}_{ijkl}) = \Psi_{ij} - \Psi_k + e_l \qquad \text{(F.14)}$$

where J^{-1} is a monotonic function, $\mathbf{D}_{ijkl}$ is the judged "difference" in heaviness, Ψ_{ij} is the heaviness of size-weight block of size i and weight j (with 21 levels), Ψ_k is the heaviness of the comparison block in the other hand (with four levels), and e_l is an additive effect of hand position and replication (with four levels).

The rescaled judgments were assessed by analysis of variance. Because the rescaling was based only on the subtractive model of "difference" judgment, it was neutral with respect to the size-weight illusion. Therefore, the scale-free estimates of Ψ_{ij} can be tested for additivity. Figure 17.31 shows the scale-free values of Ψ_{ij} that were derived from the rescaling. They show a similar divergent interaction to that in the simplified presentation of Fig. 17.29.

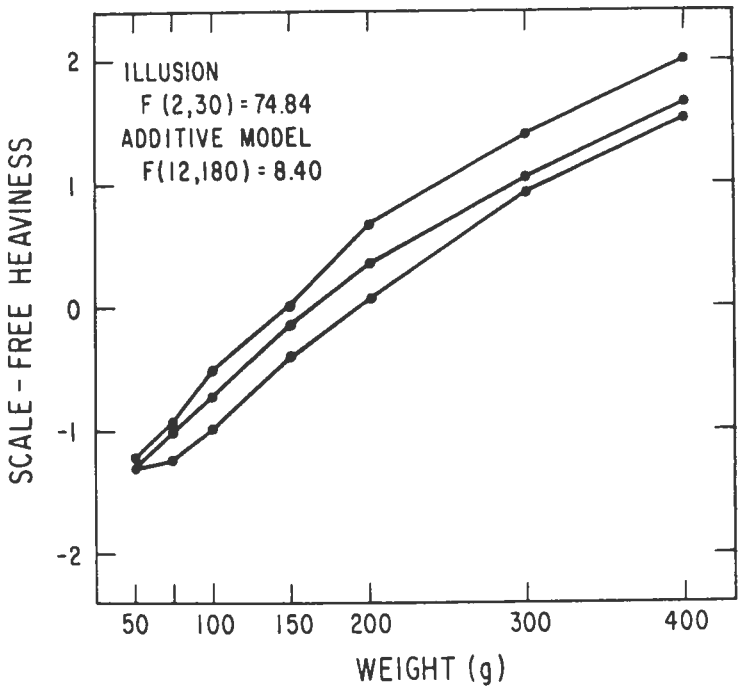


FIG. 17.31. Estimated scale-free values of heaviness from Birnbaum and Veit's (1974b) study of the size-weight illusion. Nonparallelism (divergence) implies that the additive (or parallel-averaging) model is inconsistent with the data, and cannot be salvaged by strictly monotonic transformation.

In sum, the scale-free test refutes the additive or parallel-averaging model for the size-weight illusion and impression formation, contrary to the conclusions of Anderson (1962, 1972, 1977, 1979). These findings not only require different models for these two issues, but they also show that the methodology on which the previous conclusions were based was incomplete. Implications of the size-weight interaction for heaviness perception are discussed by Birnbaum (1975) and Birnbaum and Veit (1974b).

Scale-Free Test of Ratio-Difference Theories

The scale free test (as in Fig. 17.29) is also applicable to the ratio-difference controversy (Section C). If subjects perform differences between ratios when so instructed, the data should resemble those in the left of Fig. 17.29, that is $s_7/s_1 - s_5/s_1 > s_7/s_4 - s_5/s_4$ because:

$$\frac{1}{s_1}(s_7 - s_5) > \frac{1}{s_4}(s_7 - s_5) \tag{F.15}$$

On the other hand, if subjects judge differences between differences when instructed to judge "differences between ratios," the two judgments should be equal, for:

$$(s_7 - s_1) - (s_5 - s_1) = (s_7 - s_4) - (s_5 - s_4) \tag{F.16}$$

Thus, if observers made implicit magnitude estimations of "ratios" and then computed differences, the (divergent) inequality of Expression F.15 should hold. On the other hand, if observers judge differences between differences, the equality (F.16) should hold. For the darkness experiment described in Section C, it was found that the corresponding differences were nearly equal for both the "difference of differences" and "difference of ratios" tasks. This finding is consistent with the subtractive theory.

The ratio of differences model predicts the opposite ordering (for these pairs) from that predicted by the difference of ratios model:

$$\frac{s_7 - s_1}{s_5 - s_1} < \frac{s_7 - s_4}{s_5 - s_4} \tag{F.17}$$

The data for the "ratio of differences" task showed the rank order predicted by the ratio of differences model. Surprisingly, the "ratio of ratios" task also showed a small trend in the direction predicted by the ratio of differences model. It would be useful to see further applications of the scale-free test to the ratio-difference controversy.

On "Two-Stage" Integration

Anderson (1977, 1979) recently argued that when an experiment involves three or more variables, one can "validate" the response scale by finding nonsignifi-

cant interactions between two variables and therefore trust that significant interactions among the other variables are "real" and not attributable to nonlinearity in the judgment function.

To illustrate this idea, Anderson cited a paper by Lampel and Anderson (1968) in which college women were asked to evaluate hypothetical dates based on a photograph and two personality traits. The data are shown in Fig. 17.32, plotted as a function of the level of the first adjective with a separate curve for each level of the second. Data for low, medium, and high in attractiveness of photos are shown separately.

The constant-weight averaging model predicts that the two-by-two plots for each photo should be parallel. The interaction between adjectives was nonsignificant, averaged over levels of photos (although a divergent interaction between adjectives appears when the photo is high in attractiveness). Anderson concluded that the supposed lack of adjective-by-adjective interaction "validated" the response scale, and therefore the interaction between photos and personality traits was deemed to be "real" and not an artifact of the response scale.

However, this line of argument is not consistent. If it is to be *assumed* that adjectives do not interact, it must be shown that no transformation exists that eliminates the photo-by-personality interaction and simultaneously preserves (or produces) adjective-by-adjective parallelism. By this criterion for "validity," any transformation of the data that yields adjective-by-adjective parallelism would be deemed a "valid" rescaling. Therefore, if a transformation can be

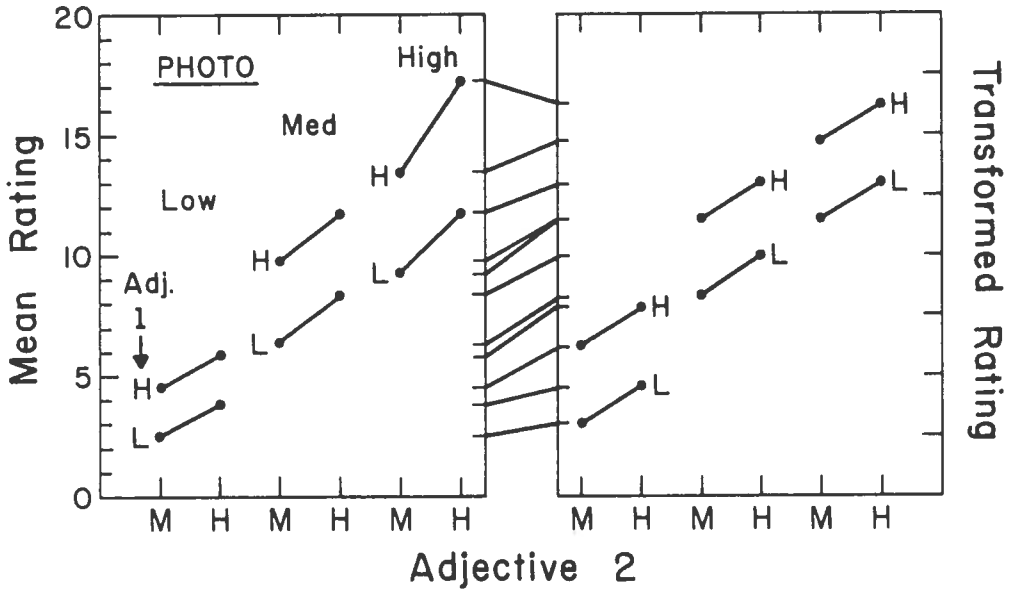


FIG. 17.32. Data of Lampel and Anderson (1968) are on the left, and rescaled data are on the right. Note that the monotonic transformation removes both the adjective-by-adjective interaction and the photo-by-adjective interaction. Therefore, these data do *not* provide convincing evidence that different operations or stages were involved. From Birnbaum (1980b).

found that eliminates the photo by adjective interactions and retains parallelism for the adjective-by-adjective interaction, one cannot use the lack of interaction between the adjectives in the raw data as evidence that the photo-by-adjective interactions are "real."

Indeed, such a monotonic transformation exists, as shown in the right of Fig. 17.32. Note that after transformation, all three adjective-by-adjective interactions are parallel (which would be deemed an "improvement" over the raw data). Furthermore, and most important, all three sets of transformed curves are congruent (differ by an additive translation), showing that the photo-by-adjective interactions have been simultaneously removed. This analysis shows that lack of significant interaction between two variables does not "validate" the "reality" of a significant interaction. The data are consistent with the hypothesis that the adjectives combine with each other by the same process as they combine with the photo. Therefore, there is no evidence for two operations in these data. (It should be mentioned that the photo-alone data, which are not presented here, would be predicted to cross the other curves according to either averaging theory under discussion and are therefore not diagnostic as to the transformation.)

These results call into serious doubt Anderson's (1974b) interpretation that the process of combining this information proceeded by "stages" in which the adjectives were first integrated by one operation and then combined with the photos by a second operation. Such a stage theory is certainly possible, but nothing in the data of Lampel and Anderson (1968) requires such an interpretation.

This discussion is not presented to argue that the divergent photo-by-adjective interaction is not "real." The assumption that adjectives do not interact is itself highly doubtful. By analogy with the results of Birnbaum (1974a), it seems likely that if a proper scale-free experiment was performed, the photo-by-adjective interaction would be confirmed. The purpose of this methodological discussion is to show that lack of significant interaction does not "validate" the response scale.

It may seem puzzling that Lampel and Anderson (1968) failed to find an adjective-by-adjective interaction. However, the dashed box in the left of Fig. 17.26 shows that the domain of their experiment was small in comparison with the domain of Birnbaum's (1974a) experiments. Apparently, the results of Birnbaum (1974a) would imply a small interaction for the small domain studied by Lampel and Anderson (1968).

A Tale of Two Ancient Philosophers

This comparison of research ranges calls to mind the tale of the two ancient Greek philosophers arguing over the shape of the earth. One of them made careful observations of plumb-bob lines separated by 100 paces. There was no evidence that the plumb-bob lines were not parallel, as they seemed to point to the same star at the same time. This parallelism was taken as evidence that the

world was flat and that the star was very far away. The other scientist travelled partway across the world and discovered that plumb-bob lines do *not* point to the same stars on the same occasions. Rather, the farther apart two plumb-bob lines are, the more they appear to diverge. He concluded that the earth is spherical, that the stars are very far away, and used his measurements to estimate the size of the earth.

During the argument some passers-by provided their suggestions: "Perhaps you are both right, and the earth changes shape depending on how one does the experiment." "Perhaps you are both wrong, and the stars are not very far away." Today, we feel both flat and spherical models deserve credit—but we do not suppose that the earth changes shape or that the stars are near. The flat-earth model has proven very useful for local surveying in small regions, and the spherical model has proven useful for navigation. Both models have since given way to more refined models of the earth. More general theories encompass early approximations, rather than refute them.

To return to the problem of impression formation, it seems reasonable to suppose that Anderson's failure to detect or take seriously the divergent interaction resulted from his use of a small variation in absolute range (difference between highest and lowest scale values in a trial). For example, a 2^6 design has 64 cells, yet only two cells have a zero within-set range (111111 and 222222). All other cells in the experiment have the *same* within-set range. If within-set range is an important determinant of the interaction, it should be clear that designs employing a small number of levels are not optimal for detecting deviations. Also, Anderson may have selected response procedures that tend to produce parallel data, thereby choosing the model in advance and finding the rescaling or response scale to agree with that model. Finally, it should be mentioned that Anderson's data do show divergent interactions.

Anderson has argued that the apparent parallelism for adjectives helped validate the rating scale and thereby form part of the argument concerning the "bias" in magnitude estimation. However, it should now be apparent that the skeptic can disregard the experiment of Anderson (1962) as a nondiagnostic study that lacked power to reveal the nonparallelism. It is therefore irrelevant to the issue of the linearity of category ratings.

Manipulation of the Interaction

Birnbaum, Wong, and Wong (1976) found divergent interactions for impression formation and also for a related task in which the judge was asked to evaluate used cars. In this study, if one adjective is bad, a person receives a low rating and the other adjectives have less effect. Similarly, if one estimate of a car's value is low, the other estimates have less effect on a buyer's opinion. Birnbaum and Stegner (1979) manipulated the interaction and actually reversed it from divergent to convergent by instructing the judges to identify with either the buyer or seller of the car. When asked to take the seller's point of view (judge the lowest

acceptable selling price), the *higher* estimate appeared to have greater weight. When asked to take the buyer's point of view (judge the highest price the buyer should pay), the *lower* estimate appeared to have greater weight.

A portion of the data from Birnbaum and Stegner's (1979) Experiment 5 is shown in Fig. 17.33. Judgments of the value of cars are plotted against the

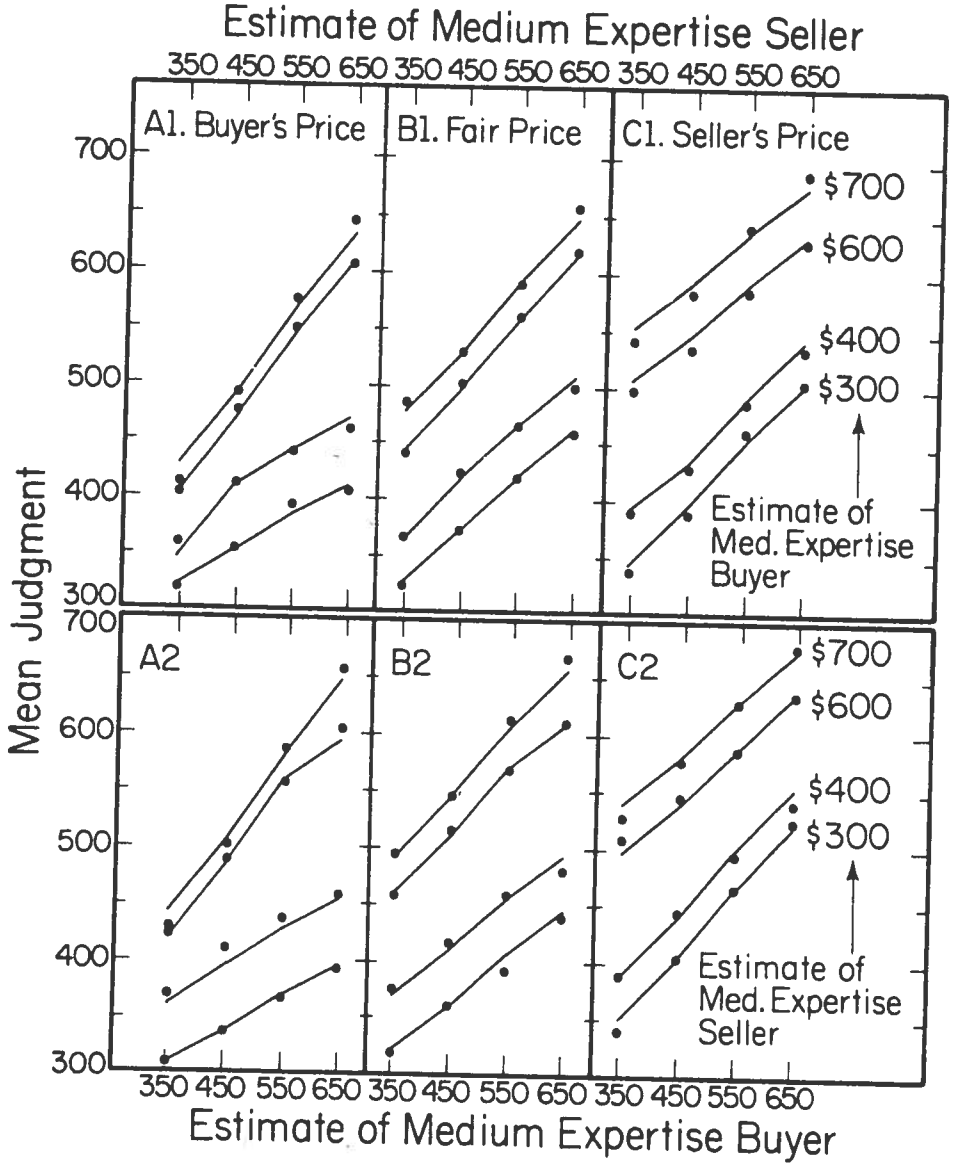


FIG. 17.33. Mean judgment of the value of used cars as a function of estimates from two mechanics who examined the cars. Panels on the left show judgments of the "highest price the buyer should pay"; panels on the right are for "lowest price seller should accept"; center panel shows judgments of "fair price." Note that curves diverge in left and center panels, and they converge in right panels. However, one cannot attribute the nonparallelism or change in shape of the curves to changes in the judgment function, for the rank order changes systematically across panels, as predicted by the configural-weight model (lines). From Birnbaum and Stegner (1979).

estimate from one source, with a separate curve for each estimate from the other source. Buyer's price and fair-price judgments show a divergent interaction, whereas seller's price judgments show a convergent interaction.

Can the results of Birnbaum and Stegner (1978, Exp. 5) be explained by assuming that the judgment function, J, depends on the judges' point of view? Recall that such an interpretation was compatible with the changing interaction in the grading on the curve experiment described in Section D (Figs. 17.19, 17.20, and 17.21). The judgment function cannot explain the change in interaction in Fig. 17.33. Note that the rank order of the points is systematically *different* in different panels. For example, from the buyer's point of view, a car with estimates of $450 and $400 is judged higher than a car with estimates of $650 and $300, whereas from the seller's point of view, the $650 and $300 car is judged about $100 higher than the $450 and $400 car. Because the rank order changes systematically, the effects of points of view *cannot* be explained in terms of a change in J. A simple configural-weight model provides a good description, using a single parameter to represent point of view, as shown by the lines representing predictions of this model (see Birnbaum & Stegner, 1979).

In order to claim understanding, one must be able to identify variables that control the effect to be explained. Thus, the ability to manipulate and even reverse the interaction by Birnbaum and Stegner (1979) goes a long way toward clarifying the process of information integration.

G. A BRIEF DISCUSSION OF RELATED THEORIES

Throughout this chapter, an attempt has been made to compare present developments with the theories of others. Parducci's range-frequency theory, discussed in Section A, was extended and elaborated in Section D. Schneider's work on the ratio-difference question is compatible with the review of that issue in Section B. Anderson's theories of the size-weight illusion, impression formation, and functional measurement are dealt with in Section F.

Several other theoretical issues, however, deserve further discussion. This section briefly reviews theories of Rule and Curtis, Eisler, Montgomery, and Marks. Rule and Curtis have a theory of estimation of "magnitudes" and "differences" that has some points of agreement and some points of disagreement with the present theory. Eisler and Montgomery have treated the relationship between category judgment and magnitude estimation in terms of the variability of the judgments on the two scales. As Montgomery has noted, their findings can be represented in terms of a single underlying scale of sensation, in agreement with the present approach. Eisler and Marks proposed that psychophysical scales change from situation to situation, depending on the subject's task. This section shows that evidence cited to argue for changes in psychophysical scales can be explained by theories that retain the premise of scale convergence.

Magnitude-Estimation Theory of Rule and Curtis

Rule and Curtis (this volume) assume that "differences" can be represented by subtraction. The heart of their theory can be stated as follows:

$$\mathbf{M}_i = J_{\mathrm{M}}(s_i) \tag{G.1}$$

$$\mathbf{MD}_{ij} = J_{\mathrm{M}}(s_i - s_j) \tag{G.2}$$

where $\mathbf{M}_i$ and $\mathbf{MD}_{ij}$ are magnitude estimations of single stimuli and of "differences," and $J_{\mathbf{M}}$ is the output function for magnitude estimation. In keeping with Attneave's (1962) theory, the $J_{\mathbf{M}}$ function, which Rule and Curtis acknowledge to depend on situational factors and individual differences, is assumed to represent (on the average) the inverse of the psychophysical function for numbers.

Their theory, in agreement with stimulus scale convergence, assumes that the psychophysical function, $s = H(\Phi)$, is independent of the task to judge "differences" or "magnitudes" and that the output function is similarly independent of these tasks. Approximating the functions H and J by power functions, Rule and Curtis represented their data as follows:

$$\mathbf{M}_i = a_{\mathbf{M}}\Phi_i^{km} + b_{\mathbf{M}} \tag{G.3}$$

$$\mathbf{MD}_{ij} = a_{\mathbf{D}}(\Phi_i^k - \Phi_j^k)^m + b_{\mathbf{D}} \tag{G.4}$$

Rule and Curtis (this volume) have estimated m from Eq. G.4 and found values between 1.1 and 2.1 with an average value of 1.47. Exponents from magnitude estimation of single stimuli are found to be close to 1.47 times larger than exponents for k derived from Eq. G.4, consistent with Eq. G.3. Furthermore, Rule and Curtis (1973) have observed that exponents for number (estimated as an input function) are close to the reciprocal of 1.47, consistent with the theory of Attneave that magnitude estimation represents cross-modality matching of numbers to stimuli and that subjective value of number is a negatively accelerated function of objective number.

In their assumptions that "differences" can be represented by subtraction, that $J_{\mathbf{M}}$ is positively accelerated, and that H is independent of the task, Rule and Curtis are in agreement with the present approach. However, they (Rule & Curtis, 1980) challenged the conclusions of Veit (1978) that subtraction can be used to represent both "ratios" and "differences" of the darkness of papers that vary in reflectance. They noted that actual ratios and differences can be monotonically related in a small, finite factorial design with a small ratio of the largest to smallest scale values. Birnbaum (1980a) has shown however, that an extension of the Rule and Curtis theory to "ratio" and "difference" judgments could not account for the data of nine experiments, using an exponent of magnitude estimations of "ratios" in the range of values cited by Rule and Curtis (this volume).

As shown in the right of Fig. 17.4, even with $m = 1.47$, $(s_j/s_j)^m$ is not monotonically related to $s_j - s_i$ in an evenly spaced seven-by-seven design with largest "ratio" judgment of 7. Birnbaum (1980a) found that for two operations to characterize the data in Fig. 17.5, one would have to reject the assumed invariance of m and use large values of m for which the power function approximates the exponential. Further comparisons and contrasts with the views of Rule and Curtis are given by Birnbaum (1980a) and Veit (1980).

Eisler's Transformation Theories

In response to the analyses and conclusions of Birnbaum (1978), reviewed in Section C, Eisler (1978) presented two theories that would allow one to retain the ratio model for judgments of "ratios." These theories, like the stage theory of Marks (1979; this volume), assume internal transformations of scales depending on the instructions given the subject.

There are two versions of Eisler's (1978) transformation theories. In one version, the subject uses only a single operation (as in the indeterminacy theory in Table 17.1) but can apply a nonlinear transformation to the scale values after this operation for "differences." In the other version, the subject uses *two* operations and a transformation that *precedes* the "difference" operation. When the ratio model is used to represent the operation for judgments of "ratios," both theories would use the logarithmic function for the transformation (T) and would predict the following:

Task	*Model*	
R:	A/B	(G.5)
D:	$T(A)-T(B)$	(G.6)
RR:	$A/B/C/D$	(G.7)
RD:	$\dfrac{T(A)-T(B)}{T(C)-T(D)}$	(G.8)
DR:	$[T(A)-T(B)] - [T(C)-T(D)]$	(G.9)
DD:	$T\left[\dfrac{T(A)-T(B)}{T(C)-T(D)}\right]$	(G.10)

In each case, the response is assumed to be a linear function of the value listed under model.

Note that the theory predicts that **DD** and **DR** should have *different* rank orders and that **DD** and **RD** should have the *same* order, contrary to the data. To handle this problem, Eisler (1978) suggested that because $[T(A)-T(B)]/[T(C)-T(D)]$ could be negative, subjects "reinterpret" the **DD** task in order to avoid the problem of negative arguments for the log. According to Eisler (1978), evidence for such a "reinterpretation" might be found by comparing the standard deviations for the

DR and **DD** tasks. However, no systematic difference appeared for the Hagerty and Birnbaum (1978) data (Birnbaum, 1979). Furthermore, in the experiment on darkness reviewed in Section C, subjects in the **DD** and **DR** tasks were instructed to compare the darker of the (A, B) pair to the lighter. This procedure guarantees that for the experimental design used, the ratio of differences (Expression G.10) for **DD** will always be $\geqslant 0$, thereby eliminating part of Eisler's (1978) rationale for the "reinterpretation" of **DD**.

Eisler (1978) remarked that Birnbaum's (1978) theory makes use of transformations (for the judgment function). He argues that the logarithmic internal transformation in his theory is as complex as the exponential output function for "ratios" in Birnbaum's theory. However, the judgment functions in Birnbaum's theory do not affect the rank order of the data, so at the ordinal level the judgment functions can be disregarded, whereas nonlinear internal transformations will alter rank order. Furthermore, Birnbaum and Veit's (1974a) theory of the judgment function predicts that the range of examples will affect the response range. Finally, one has to acknowledge judgment functions to account for the changes that can be induced by means of the context (Sections A and D). Additional comments on various details of Eisler's (1978) transformation theories are given in Birnbaum (1979).

In summary, the transformation theories seem unattractive, not only because of the postulated internal transformation, but also because of the "reinterpretation" argument that must be made in order to rectify an otherwise incorrect prediction of the theory, and thereby make the theory ordinally equivalent to the subtractive theory.

Stage Theory of Marks

Marks (1979) proposed a stage theory of loudness that has some similarities to Eisler's theory. The essentials of his theory can be summarized as follows:

$$\mathbf{M}_i = J_{\mathbf{M}}[L_i] \tag{G.11}$$

$$\mathbf{MT}_{ij} = J_{\mathbf{T}}[L_i + L_j] \tag{G.12}$$

$$\mathbf{MD}_{ij} = J_{\mathbf{D}}[T(L_i) - T(L_j)] \tag{G.13}$$

$$\mathbf{MDT}_{ijkl} = J_{\mathbf{D}}[T(L_i + L_j) - T(L_k + L_l)] \tag{G.14}$$

where $\mathbf{M}_i$, $\mathbf{MT}_{ij}$, $\mathbf{MD}_{ij}$, and $\mathbf{MDT}_{ijkl}$ are "magnitude estimations" of the single stimuli, magnitude estimations of "total loudness" of multicomponent (or binaurally presented) tones, magnitude estimations of "differences" between two tones, and magnitude estimations of the "difference" between two "summated" loudnesses, respectively. L_i is the subjective loudness of a tone or component, $J_{\mathbf{M}}$, $J_{\mathbf{T}}$, and $J_{\mathbf{D}}$ are monotonic output (judgment) functions for magnitude estimation in these tasks, and T represents the transformation between the so-called "L"

scale and what Marks (1979) terms the "D" scale of loudness, where $D = T(L)$. In Marks (1979) paper, T is approximated by the square-root function. (Eisler [1978] would represent T with the log function.) Marks argues that by averaging exponents across different experiments, the J_{M} function can be assumed to be a similarity function. However, he concludes that the J_{T} function is negatively accelerated (Marks approximates it by using a power function with exponent of $\frac{3}{4}$), and he presents evidence that the J_{D} function is positively accelerated for magnitude estimations of "differences." This agrees with Rule and Curtis (this volume), who approximate J_{D} by a power function with an average exponent of about 1.5.

The theory of Marks (1979) represents binaural and multicomponent summation by the arithmetic addition operation; it represents "difference" (and perhaps also "ratios") by the subtraction operation. However, in order to do this, three systematically different output transformations are required for magnitude estimations, and two different input transformations are used for loudness. (Actually, the number of transformations and output functions is still larger when one considers Marks' treatment of loudness addition within the critical band width, which is not treated here.)

Marks (1979) theory seems unduly complicated to account for the data he reviews. First, as is shown later, it is possible to retain a single scale of loudness for both "differences" and loudness "summation." Therefore, the T transformation appears unnecessary. Second, it can be shown that it is possible to explain the data using a single J.

A simpler theory that preserves scale convergence can be written as follows:

$$\mathbf{M}_i = J_{\mathbf{M}}[s_i] \tag{G.15}$$

$$\mathbf{MT}_{ij} = J_{\mathbf{T}}[\Psi_{ij}] \tag{G.16}$$

$$\mathbf{MD}_{ij} = J_{\mathbf{D}}[s_i - s_j] \tag{G.17}$$

$$\mathbf{MDT}_{ijkl} = J_{\mathbf{D}}[\Psi_{ij} - \Psi_{kl}] \tag{G.18}$$

where Ψ_{ij} is the overall loudness experience produced by s_i and s_j, and $\Psi_{ij} = C(s_i, s_j)$ represents the combination function for loudness "summation." This theory preserves scale convergence at the expense of representing combination by a *nonadditive* function, as in Birnbaum et al. (1971). It also allows one to retain the same theory for all of the J functions in Eqs. G.15 through G.18.

For simplicity and comparability with the work of others who approximate data by power functions, a rough approximation could be stated as follows:

$$\Psi_{ij} = \sqrt{s_i^2 + s_j^2} \tag{G.19}$$

where Ψ_{ij} is the "total loudness" of tones having scale values s_i and s_j. Equation G.19 is the equation for the length of the sum of two orthogonal vectors. (Equation G.19 could be generalized to include the angle between the

vectors or replaced by a similar function predicting a convergent interaction to provide a more accurate representation.) Figure 17.34 shows Ψ_{ij} plotted as a function of s_i, with a separate curve for each of several values of s_j, according to Eq. G.19. Note that this theory of loudness does not predict parallelism, but predicts a convergent interaction, similar to that obtained by Marks (1979, and this volume).

It may be possible to assume that all of the output functions are governed by an exponent of about 1.5, consistent with the findings of Rule and Curtis (this volume). Note that Eqs. G.15, G.16, and G.17 become:

$$\mathbf{M}_i = a s_i^{1.5} \tag{G.20}$$

$$\mathbf{MT}_{ij} = a_{\mathbf{T}}[(s_i^2 + s_j^2)^{.5}]^{1.5} \tag{G.21}$$

$$\mathbf{MD}_{ij} = a_{\mathbf{D}}[s_i - s_j]^{1.5} + b_{\mathbf{D}} \tag{G.22}$$

It follows that under Marks' (1979) analysis, the output function for loudness summation would be .75, so the $\frac{4}{3}$ exponent empirically obtained by Marks (1979) in his equation

$$\mathbf{MT}_{ij}^{4/3} = L_i + L_j \tag{G.23}$$

is *predicted* by this theory because Eq. G.21 becomes $\mathbf{MT}_{ij}^{4/3} = s_i^2 + s_j^2$.

Furthermore, the relationship between Marks' "*L*"scale and his "*D*" scale would also be *predicted* by this theory to be the square function in this approximation, consistent with the conclusion reached by Marks (1979). In Eq. G.21, s_i corresponds to D and s_i^2 to L. It should also be noted that the present theory handles "differences between summated loudnesses" without the postulated

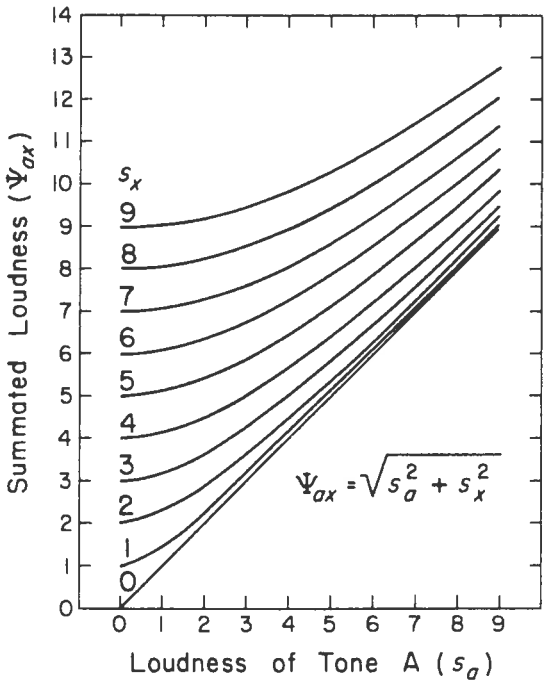


FIG. 17.34. Theoretical curves for loudness "summation" based on Pythagorean rather than arithmetic addition. Note that the difference in overall loudness due to variation of one component is less when the other component is loud than when it is soft. From Birnbaum (1980d).

internal T transformation, representing them as monotonically related to subjective differences in Ψ of Eq. G.19:

$$\mathbf{MDT}_{ijkl} = J_{\mathbf{D}}[\Psi_{ij} - \Psi_{kl}] \tag{G.24}$$

where $\mathbf{MDT}_{ijkl}$ is the magnitude estimation of the "difference" in loudness between two "summated" loudness experiences.

It is not seriously suggested that these power functions are the correct theoretical representation. They are used only for simplicity and comparability with the work of Marks. Any complete theory must be able to predict contextual effects on the J functions. However, the foregoing theory, oversimplified as it is, provides at least as good a representation of the data reviewed by Marks' (1979) as his theory (Eqs. G.11-G.14). Marks (1979) theory requires two scales (D and L) and three functions for magnitude estimation. It has *no theory to predict* the relationship between D and L and *no theory to explain* why $J_{\mathbf{T}}$ should be negatively accelerating, $J_{\mathbf{D}}$ should be positively accelerating, and $J_{\mathbf{M}}$ should be linear. The suggested theory (Eqs. G.15-G.18) uses only one scale (s_i) and one theory for the J function of magnitude estimation.

Loudness Combination

Falmagne, Iverson, and Marcovici (1979) developed a theory of loudness discrimination and combination. They presented the observer with two binaural tone pairs, (a, x) and (b, y), and asked the observer to report which pair produced the louder experience. They concluded that their data can be represented by the equations:

$$\Psi_{ax} = s_a + s_x \tag{G.25}$$

$$\mathbf{P}_{ax;by} = F[T(\Psi_{ax}) - T(\Psi_{by})] \tag{G.26}$$

where Ψ_{ax} is the overall loudness of tones a and x, s_a and s_x are the loudnesses of the tones presented to the left and right ears, $\mathbf{P}_{ax;by}$ is the proportion of responses indicating (a, x) is louder than (b, y), and F is a cumulative density function. Note that T performs a role similar to that in Marks' theory.

Falmagne et al. (1979) found that $\mathbf{P}_{ax;ay}$ decreases as a function of a, indicating either that Ψ_{ax} does not equal $s_a + s_x$ or that T is negatively accelerated. Falmagne et al. (1979) represented T as a logarithmic function (rather than as the square-root function used by Marks). This interpretation, they noted, represents choice probabilities as monotonically related to subjective ratios rather than as differences. If the psychophysical function is also assumed to be a power function, their theory implies a "conjoint Weber's law," in which choice probabilities for (a, x) vs. (b, y) should be the same as (ta, tx) vs. (tb, ty) for any value of t. It remains to be seen whether the conjoint Weber's law will prove more than a rough approximation when tested over a wider stimulus range.

The finding that $\mathbf{P}_{ax;ay}$ decreases as a function of a is consistent with the interaction shown in Fig. 17.34 and the assumption that T is an identity function.

Such an interpretation would be hard to discriminate from the theory of Falmagne et al. (1979) on the basis of their experiments.

Therefore, it may be simpler to represent loudness summation by vector addition rather than arithmetic addition. Such a representation permits one to retain scale convergence for both input (psychophysical) and output (judgmental) functions, and seems therefore simpler and greater in explanatory power than Marks' theory, which requires two scales, three output functions, and has no theory to explain their relationship. Such a representation would also predict the major result of Falmagne et al. (1979) in terms of a single scale of sensation and a subtractive theory for stimulus comparison and discrimination.

A Transformation-Theory Account of Impression Formation

In Section F, the additive (or parallel-averaging) model of impression formation was found to be inconsistent with ratings of "combined" likeableness and "differences" in likeableness if the stimulus scale convergence criterion was assumed. It seems reasonable to ask if the transformation theories of Marks and Eisler could be extended in a consistent fashion (using the same T) to encompass impression formation.

The following transformation theory can describe the results of Birnbaum (1974a):

$$\mathbf{D}_{ij} = J_{\mathbf{D}}[T^*(s_i) - T^*(\mathrm{s}_j)] \tag{G.27}$$

$$\mathbf{C}_{ij} = J_{\mathbf{C}}[s_i + s_j] \tag{G.28}$$

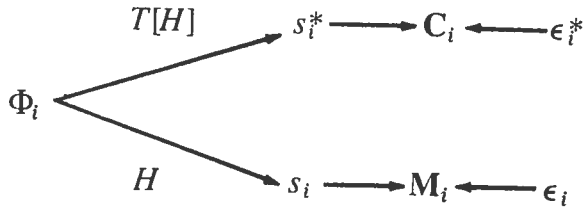
$$\mathbf{DC}_{ijkl} = J_{\mathbf{DC}}[T^*(\mathrm{s}_i + s_j) - T^*(\mathrm{s}_k + s_l)] \tag{G.29}$$

where $J_{\mathbf{D}}$ and $J_{\mathbf{DC}}$ are approximately linear, but $J_{\mathbf{C}}$ is positively accelerated, and T^* is positively accelerated. According to the stage theories of Marks and Eisler, however, all of the J functions should be negatively accelerating, and furthermore, T^* should be negatively accelerating. Thus, the stage theory cannot describe "differences" in terms of the *same* T^* transformation to salvage both the additive models of loudness "summation" and impression information. These results, therefore, put the transformation theory in the post hoc position of requiring different transformations for every situation. As shown in Section F, a coherent account of the data for impression formation can be given in terms of a single scale of likeableness for the adjectives and the assumption that all of the J functions are approximately linear.

Psychophysical Variability

Eisler (1963) has given an account of the relationship between category ratings and magnitude estimations in terms of the variability of the two scales. This approach is reviewed by Montgomery (this volume), who notes that the results can be given a different interpretation from Eisler's.

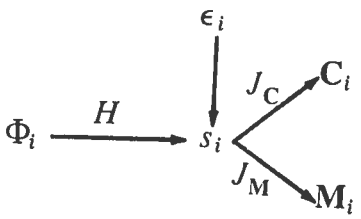
Eisler's interpretation can be diagrammed as follows:



(G.30)

where H is the psychophysical function, T is the transformation, s and s^* are the two scales of subjective value, and ϵ_i^* and ϵ_i are the error terms, representing variability. According to Eisler's early development, the variance of ϵ_i^* was assumed constant with respect to s_i^*, but the variance of ϵ_i was assumed to be linearly related to s_i. This was interpreted as a subjective Weber's law.

An alternative interpretation can be given as follows:



(G.31)

This theory (a special case of that discussed by Birnbaum, 1979) uses only one scale of sensation and one error term, but different $J_{\mathbf{C}}$ and $J_{\mathbf{M}}$ functions for category rating and magnitude estimation in different contexts. This theory leads to the general psychophysical differential equation used by Eisler and Montgomery. It predicts that when the stimuli are scaled in accordance with Thurstone's law of categorical judgment, the estimated scale values should be independent of stimulus distribution or the task to use category ratings or magnitude estimations. Parducci (1965, 1974) has shown that Thurstone scales are nearly independent of stimulus distribution. Montgomery (this volume) has shown that although the relationship between category ratings and magnitude estimations differs for different individuals, the data can be well-approximated by the theory of one scale, assuming that the variance of ϵ_i is independent of s_i, that $J_{\mathbf{C}}$ is approximately linear for most subjects (for his stimulus spacing), and that $J_{\mathbf{M}}$is nonlinear and varies for different individuals.

Birnbaum (1979) noted that Thurstone's simplest case (equal variances) could be applicable to matrices of both category ratings and magnitude estimations. Let $\mathbf{P}_{ij}$ and $\mathbf{Q}_{ik}$ be the cumulative proportion of responses to stimulus i less than or equal to category "j" or magnitude-estimation response "X_k," respectively. Then, one can write:

$$\mathbf{P}_{ij} = F[(s_i - t_j)/a] \tag{G.32}$$

$$\mathbf{Q}_{ik} = F[(s_i - u_k)/b] \tag{G.33}$$

where F is a cumulative density function (e.g., normal). If there is only a single error term, as in Expression G.31, and if the variance of ϵ_i is independent of s_i,

then a can be set equal to b equal to 1.0. The relationship between t_j and j describes $J_{\mathbf{C}}$; the relationship between u_k and X_k describes $J_{\mathbf{M}}$. In this special case (where $a = b$), it should be possible to find a spacing of magnitude estimations (X_k) for each subject, such that $\mathbf{P}_{ij} = \mathbf{Q}_{ij}$. Montgomery (this volume) has shown that this prediction may be a reasonable approximation.

In conclusion, the assumption that there is a single scale of sensation with a single source of subjective variability provides a reasonable theoretical representation of the empirical relationship between means and standard deviations of category ratings and magnitude estimations. Analyses of "differences," "ratios," and "totals" may provide further insight into the loci of psychophysical and judgmental variability.

CONCLUDING COMMENTS

This chapter offers the following resolutions to the measurement controversies:

1. Overt judgments can be regarded as a monotonic function of subjective value, where the nature of the monotonic function depends lawfully on the stimulus and response distributions. The range and frequency distribution of both stimuli and responses affects the nature of this function.

2. Judgments of "ratios" and "differences" for most continua can be represented by the subtractive model using the same scale values for both tasks:

$$\mathbf{R} = J_{\mathbf{R}}[\mathrm{A} - \mathrm{B}]$$

$$\mathbf{D} = J_{\mathbf{D}}[\mathrm{A} - \mathrm{B}]$$

3. Judgments of "ratios of ratios," "differences of ratios," "ratios of differences," and "differences of differences" can be represented by the subtractive theory:

$$\mathbf{RR} = J_{\mathbf{RR}}[(\mathrm{A} - \mathrm{B}) - (\mathrm{C} - \mathrm{D})]$$

$$\mathbf{DR} = J_{\mathbf{DR}}[(\mathrm{A} - \mathrm{B}) - (\mathrm{C} - \mathrm{D})]$$

$$\mathbf{RD} = J_{\mathbf{RD}}[(\mathrm{A} - \mathrm{B})/(\mathrm{C} - \mathrm{D})]$$

$$\mathbf{DD} = J_{\mathbf{DD}}[(\mathrm{A} - \mathrm{B}) - (\mathrm{C} - \mathrm{D})]$$

4. The judgment functions are approximately linear for category ratings and approximately exponential for magnitude estimation when the stimuli are geometrically spaced, the category-response examples are equally spaced, and the magnitude-estimation examples are geometrically spaced.

5. Scale values estimated from the subtractive theory of (within-mode) stimulus comparison appear largely independent of stimulus spacing.

6. The judgment function in information-combination experiments appears to depend on the distribution of subjective combinations.

7. Scale values estimated from cross-modality comparison and combination depend on the stimulus distribution.

8. One should neither select a standardized method for conducting experimental research on the basis of a priori considerations, nor attempt to "avoid" contextual effects by holding context fixed to some arbitrary value. Instead, it seems reasonable to manipulate procedures and contexts systematically and to base generalizations on empirically established laws of judgment.

9. The fit of a model does not simultaneously validate the response scale and model. Previous conclusions that impression formation and the size-weight illusion obey a simple averaging model (additive) were based on inappropriate conclusions from functional measurement. Methods involving scale convergence and the scale-free test should be applied to provide more strenuous tests of algebraic models.

10. Theories assuming that measurements of subjective value transcend the tasks from which they were derived should be preferred to theories assuming different scales. In particular, by representing loudness additivity with Pythagorean addition rather than arithmetic addition, it may be possible to retain scale convergence for stimulus combination and comparison. Similarly, by representing impression formation with a configural model, it is possible to retain the premise of scale convergence for combination and comparison.

Despite generations of controversy concerning the theoretical representation of subjective value and even the appropriate models and methods for measurement of subjective value, a number of empirical findings emerge that show lawful regularity. The lawfulness of stimulus comparison and combination and the regularity of contextual effects constitute results that must be explained by any viable theory. The premises just listed may provide the beginnings of a coherent solution to the controversies of psychological measurement.

ACKNOWLEDGMENTS

Preparation of this chapter was supported in part by the Research Board of the University of Illinois. Thanks are due Barbara Mellers for her valuable assistance and advice.

REFERENCES

Anderson, N. H. Application of an additive model to impression formation. *Science*, 1962, *138*, 817–818.

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

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

Anderson, N. H. Cross-task validation of functional measurement using judgments of total magnitude. *Journal of Experimental Psychology*, 1974, *102*, 226-233. (a)

Anderson, N. H. Information integration theory: A brief survey. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2). San Francisco: Freeman, 1974. (b)

Anderson, N. H. Note on functional measurement and data analysis. *Perception & Psychophysics*, 1977, *21*, 201-215.

Anderson, N. H. Algebraic rules and psychological measurement. *American Scientist*, 1979, *67*, 555-563.

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

Baird, J. C. & Noma, E. *Fundamentals of scaling and psychophysics*. New York: Wiley, 1978.

Birnbaum, M. H. Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology*, 1972, *93*, 35-42.

Birnbaum, M. H. The devil rides again: Correlation as an index of fit. *Psychological Bulletin*, 1973, *79*, 239-242.

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

Birnbaum, M. H. Reply to the devil's advocates: Don't confound model testing and measurement. *Psychological Bulletin*, 1974, *81*, 854-859. (b)

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

Birnbaum, M. H. Expectancy and judgment. In F. Restle, R. Shiffrin, N. J. Castellan, H. Lindman, & D. Pisoni (Eds.), *Cognitive theory* (Vol. 1). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

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

Birnbaum, M. H. Reply to Eisler: On the subtractive theory of stimulus comparison. *Perception & Psychophysics*, 1979, *25*, 150-156.

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

Birnbaum, M. H. *Issues in functional measurement*. Unpublished manuscript, 1980. (Available from author, Dept. of Psychology, University of Illinois, 603 E. Daniel, Champaign, IL 61820) (b)

Birnbaum, M. H. *Systextual design*. Unpublished manuscript, 1980. (Available from author) (c)

Birnbaum, M. H. *Toward a coherent theory of psychophysical judgment*. Unpublished manuscript, 1980. (Available from author) (d)

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

Birnbaum, M. H., Kobernick, M., & Veit, C. T. Subjective correlation and the size-numerosity illusion. *Journal of Experimental Psychology*, 1974, *102*, 537-539.

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

Birnbaum, M. H., & Mellers, B. A. *Context effects in category rating and magnitude estimation*. Unpublished manuscript, 1980. (Available from author) (a)

Birnbaum, M. H., & Mellers, B. A. *Context effects in cross-modality comparison and combination*. Unpublished manuscript, 1980. (Available from author) (b)

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. E. Source credibility: Expertise, bias, and the judge's point of view. *Journal of Personality and Social Psychology*, 1979, *37*, 48-74.

Birnbaum, M. H., & Veit, C. T. Judgmental illusion produced by contrast with expectancy. *Perception & Psychophysics*, 1973, *13*, 149-152.

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. (a)

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

Birnbaum, M. H., Wong, R., & Wong, L. Combining information from sources that vary in credibility. *Memory & Cognition*, 1976, *4*, 330-336.

Brunswik, E. *Perception and the representative design of experiments*. Berkeley: University of California Press, 1956.

Curtis, D. W., & Rule, S. J. Judgments of average lightness and darkness: A further consideration of inverse attributes. *Perception & Psychophysics*, 1978, *24*, 343-348.

Eisler, H. On the problem of category scales in psychophysics. *Scandinavian Journal of Psychology*, 1962, *3*, 81-87.

Eisler, H. A general differential equation in psychophysics: Derivation and empirical test. *Scandinavian Journal of Psychology*, 1963, *4*, 265-272.

Eisler, H. On the ability to estimate differences: A note on Birnbaum's subtractive model. *Perception & Psychophysics*, 1978, *24*, 185-189.

Elmasian, R., & Birnbaum, M. H. *A harmonious note on pitch*. Unpublished manuscript, 1979. (Available from author)

Falmagne, J. C., Iverson, G., & Marcovici, S. Binaural "loudness" summation: Probabilistic theory and data. *Psychological Review*, 1979, *86*, 25-43.

Feldman, J., & Baird, J. C. Magnitude estimation of multidimensional stimuli. *Perception & Psychophysics*, 1971, *10*, 418-422.

Hagerty, M., & Birnbaum, M. H. Nonmetric tests of ratio vs. subtractive theories of stimulus comparison. *Perception & Psychophysics*, 1978, *24*, 121-129.

Jones, C., & Aronson, E. Attribution of fault to a rape victim as a function of the respectability of the victim. *Journal of Personality and Social Psychology*, 1973, *26*, 415-419.

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

Krantz, D. H., Luce, R. D., Suppes, D., & 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.

Krumhansl, C. L. Concerning the applicability of geometric models of similarity to data: The interrelationship between similarity and density. *Psychological Review*, 1978, *85*, 445-463.

Lampel, A. K., & Anderson, N. H. Combining visual and verbal information in an impression-formation task. *Journal of Personality and Social Psychology*, 1968, *9*, 1-6.

Luce, R. D. What sort of measurement is psychophysical measurement? *American Psychologist*, 1972, *27*, 96-106.

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

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

Mellers, B. A., & Birnbaum, M. H. Context effects in social judgment. Unpublished manuscript, 1980. (Available from author) (a)

Mellers, B. A., & Birnbaum, M. H. Context effects in within-mode stimulus comparison. Unpublished manuscript, 1980. (Available from author) (b)

Nihm, S. D. Polynomial law of sensation. *American Psychologist*, 1976, *31*, 808-809.

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

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

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

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

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

Petrinovich, L. Probabilistic functionalism: A conception of research method. *American Psychologist*, 1979, *34*, 373-390.

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

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

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

Restle, F. A metric and an ordering on sets. *Psychometrika*, 1959, *24*, 207-220.

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

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.

Rule, S. J., & Curtis, D. W. Ordinal properties of subjective ratios and differences. *Journal of Experimental Psychology: General*, 1980, *109*, 296-300.

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.

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

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

Sidowski, J. B., & Anderson, N. H. Judgments of city-occupation combinations. *Psychonomic Science*, 1967, *7*, 279-280.

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

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

Teghtsoonian, R. On the exponents in 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.

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

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

Veit, C. T. Analyzing "ratio" and "difference" judgments: A reply to Rule and Curtis. *Journal of Experimental Psychology: General*, 1980, *109*, 301-303.

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