Perception & Psychophysics
1974, Vol. 15, No. 1, 7-15

Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks*

MICHAEL H. BIRNBAUM†
University of California, San Diego, La Jolla, California 92037

and

CLAIRICE T. VEIT
University of California, Los Angeles, Los Angeles, California 90024

Ss lifted pairs of weights simultaneously, one in each hand, and judged either the difference, ratio, or average heaviness of the two weights. Data for the difference and ratio tasks were in general agreement with subtractive and ratio models, but the averaging data showed discrepancies from the constant-weight averaging model similar to those reported in previous psychophysical research. Rescaling was ruled out for the averaging data, because responses to pairs of equal weight were a linear function of subtractive model scale values derived from the difference task data. Scale values for the ratio and difference task data were related exponentially, as were the responses to the pairs, consistent with Torgerson's conjecture that Ss do not distinguish "differences" from "ratios." They appear to use the same composition rule but different output functions, depending on the procedures for responding. The scale convergence criterion can thus prevent inappropriate rescaling when a model fails and can dictate rescaling even when a model fits.

Traditionally, psychophysics was defined as the study of the relationships between physical stimuli and subjective impressions. A popular, contemporary type of psychophysical scaling, such as that advocated by Stevens (1957, 1971), can be represented as in Fig. 1A. In this schema, H represents the psychophysical function relating physical values (ϕ) to impressions (s), and J represents the function relating overt responses, R, to the impressions. Such "direct" measurement requires untested assumptions (Treisman, 1964; Savage, 1966). For example, to obtain a scale of sensation from magnitude estimations of single stimuli, one must assume that the J function relating magnitude estimations to subjective values is linear, with a zero intercept. Similarly, one might derive a scale of sensation from category ratings through the analogous assumption that ratings involve a linear transformation, J.

If both procedures were valid, magnitude estimation and category judgment scales would be linearly related; instead, magnitude estimations are usually a positively accelerated function of category judgments (Stevens & Galanter, 1957). This empirical contradiction throws doubt on any conclusions concerning the form of H or J.

With reference to the outline in Fig. 1A, magnitude estimation data could be explained by postulating any pair of functions whose composition is a power function: for example, H power and J power, H identity and J power, H linear-logarithmic and J exponential. Unifactor methodology does not provide enough leverage to discriminate between these alternatives.

*We thank Allen Parducci and Norman H. Anderson for their helpful comments on earlier versions of this paper. Computing assistance was received from Campus Computing Network, University of California, Los Angeles. The first author received support from a National Institute of Mental Health postdoctoral fellowship at University of California, San Diego. Additional support was provided by the Center for Human Information Processing, through NIMH Grant MH-15828, and by NSF Grant GB-21028.

†Requests for reprints should be sent to Michael H. Birnbaum, Department of Psychology, Kansas State University, Manhattan, Kansas 66506.

Information Integration

A concern for the logic of measurement has redirected the focus from untestable, theoryless scaling to an increased study of algebraic models that permit simultaneous evaluation of a postulated theory and scaling of the stimuli (Anderson, 1970; Cliff, 1973; Krantz, 1972; Krantz, Luce, Suppes, & Tversky, 1971; Zinnes, 1969). This approach is represented in Fig. 1B. The basic idea is that by considering the contribution of two or more factors simultaneously, it is possible to measure each factor separately; the resultant scale values are based upon a theory of integration and have meaning with respect to the theory.

The conjoint scaling approach, related to conjoint measurement (Krantz et al, 1971), assumes a valid composition rule, I. If the data are ordinally inconsistent with the composition rule, then the model for I can be rejected (Krantz & Tversky, 1971). Otherwise, the data are transformed to fit the assumed model. Only the ordinal information in the raw data is required in order to derive scales compatible with the assumed model of integration. This approach is not entirely satisfactory, however, for if J is linear, then nonlinear monotone rescaling might remove true discrepancies from the model (Birnbaum, 1972).

It is not possible to distinguish between models that are monotone transforms of one another when rescaling is permitted. In this case, scales become arbitrarily

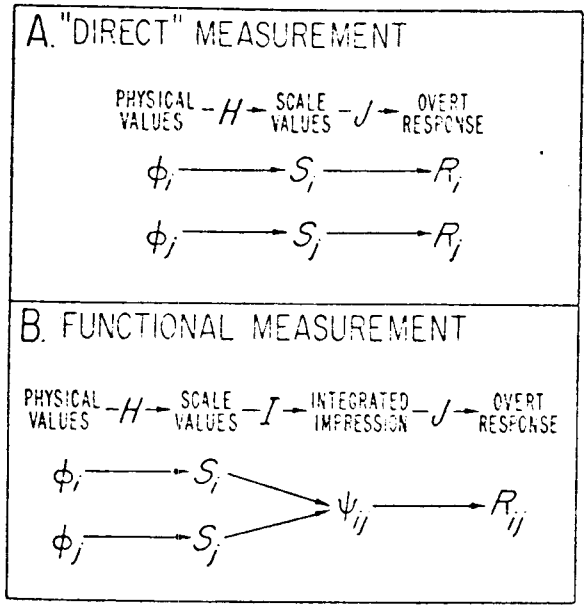

Fig. 1. (A) Outline of "direct" measurement. H represents the psychophysical function that transforms the physical value of the stimulus, ϕ into a psychological scale value, s. J represents the judgment function which transforms the s value into an overt response, R. (B) Outline of functional measurement. The physical values of the component stimuli, ϕ_i, are transformed to psychological scale values, s_i, by the psychophysical function, H. These scale values are combined by the integration rule, I, to form an overall impression, Ψ_{ij}, which is then transformed into an overt response, R_{ij}, by the judgment function, J.

dependent upon the assumed, but untested, model. For example, since a logarithmic transformation of a ratio yields a difference, this approach cannot discriminate between subtractive and ratio models (Krantz et al, 1971).

Functional measurement (Anderson, 1970) attempts to simultaneously evaluate the composition rule and the response function. With the J function fixed, difference and ratio models make different predictions for the raw data. If the raw data satisfy tests of fit, then the validity of the scale values, the integration model, and response scale are jointly supported. Otherwise, either the model or response scale (or both) are invalidated.

Generally, functional measurement discriminates between model and response scale discrepancies in the following manner: If there is no reason to doubt the validity of the response scale, the test of fit provides the basis for rejecting the model; if there is reason to believe that the response scale is nonlinear (e.g., when the response is on the physical continuum, as in the method of bisection), a search is made for a monotone transformation that will fit the model (Anderson, 1962, 1970; Bogartz & Wackwitz, 1971). The decision to rescale the data or to reject the model is based upon subtle theoretical considerations such as the nature of the necessary transformation, the history of previous experiments with the model and response mode, signs of end or ceiling effects, and their dependence upon variations in experimental procedures.

Scale Convergence Criterion

The following additional constraint was proposed (Birnbaum, 1972) to aid in determining whether deviations of fit from the model are due to an inappropriate model or an inappropriate response scale: the scale values of the stimuli are assumed to be independent of the composition rule. The psychophysical function, H, which relates scale values derived from the hypothesized model to the physical values of the stimuli, is thus constrained to be independent of the integration task.

That measurements of the same stimuli by different techniques should agree is one of the most obvious and appealing concepts in psychology (Seward, 1955; Garner, Hake, & Eriksen, 1956; Anderson, 1962; Krantz, 1972; Cliff, 1973). In the past, convergence of operations (cf. Stevens & Galanter, 1957) and the fit of models (cf. Anderson, 1970) have been investigated separately, but they have less often been studied together.

As stated above, it is not always appropriate to transform otherwise inconsistent data to fit an assumed model. The scale convergence criterion, however, suggests that an *appropriate* transformation be defined as one that both makes the model fit and leads to the derivation of scale values that are appropriately related to those derived from the fit of another model in another situation.

In the present experiment, Ss are required to lift two weights simultaneously in the two hands and judge either (a) the *difference* in heaviness between the weights in the two hands, (b) the *ratio* of the heavinesses, or (c) the *average* heaviness of the two weights. A different model is considered for each task, but the scale values of the stimuli are assumed to be the same, independent of task. The constraint is sufficient to provide a basis for response transformation that is independent of the model under investigation.

The Models

The psychological differences are assumed to follow a subtractive model:

$$\Psi^{D}_{ij} = s_i - s_j, \tag{1}$$

where s_i and s_j are the scale values of the i^{th} and j^{th} levels of the stimuli presented to the right and left hands, and Ψ^{D}_{ij} is the psychological value of the difference between the stimuli in the two hands.

The psychological ratios, Ψ^{R}_{ij}, are assumed to follow a ratio model:

$$\Psi^{R}_{ij} = s_i/s_j. \tag{2}$$

The averaging model assumes that the psychological average, Ψ^{A}_{ij}, is a weighted average of the scale values:

$$\Psi_{ij}^{A} = ws_i + (1 - w)s_j, \quad (3)$$

where w and 1 − w are the relative importances of the stimuli in either hand.

The overt responses for each task are initially assumed to be linear functions of the psychological impressions:

$$R = a\Psi + b, \quad (4)$$

where a and b would depend upon the task and the response mode. This assumption is evaluated with respect to two criteria: (a) fit of the hypothesized model to the raw data, and (b) convergence of scale values between tasks.

METHOD

Ss were run individually, seated at a table separated by a screen from E. On each trial, E placed two stimulus weights before S, who lifted them simultaneously and judged the difference, the ratio, or the average of the heavinesses of the two stimuli.

Stimuli and Design

The stimuli were plastic cylinders, approximately 8 cm tall and 3 cm in diam. The cylinders were filled with cotton and lead shot and were lined with opaque paper so that they were all identical in appearance.

The stimulus pairs represent a symmetric 7 by 7 factorial design, in which the weight of the stimulus in either hand could be 50, 75, 100, 125, 150, 175, or 200 g.

There were six different random sequences of the 49 paired presentations, which were randomly counterbalanced for the task: each S used a different sequence for each task, but an equal number of Ss in each task used each sequence.

Each of 24 Ss performed all three tasks; 4 Ss performed the tasks in each of the six possible orders. After the instructions were read for each task, nine representative warm-up trials were given to familiarize S with the new task and response mode, and to decrease transfer between tasks. Since there was no discernible effect of task order or sequence order for the data in any of the tasks, order is ignored in the subsequent data analyses.

Instructions

Difference Task

Ss rated the subjective difference in heaviness on a 9-point scale, from 1 (right hand is *very very much lighter* than the left) to 9 (right hand is *very very much heavier* than the left).

Ratio Task

Ss were instructed to report the ratio of heaviness of the weight in the right hand to the heaviness of the weight in the left hand. The instructions permitted the use of any numbers to express this ratio, but Ss were given a page with the following printed examples: 25 = right hand is *one-fourth as heavy* as the left, 33 = right hand is *one-third as heavy* as the left, 50 = right hand is *one-half as heavy* as the left, 100 = ratio equals *one*, 200 = right hand is *twice as heavy* as the left, 300 = right hand is *three times as heavy* as the left, and 400 = right hand is *four times as heavy* as the left.

Averaging Task

Ss were instructed to "judge the average heaviness of the weights in the two hands." They used a 9-category scale with labels varying from 1 (*very very light*) to 9 (*very very heavy*).

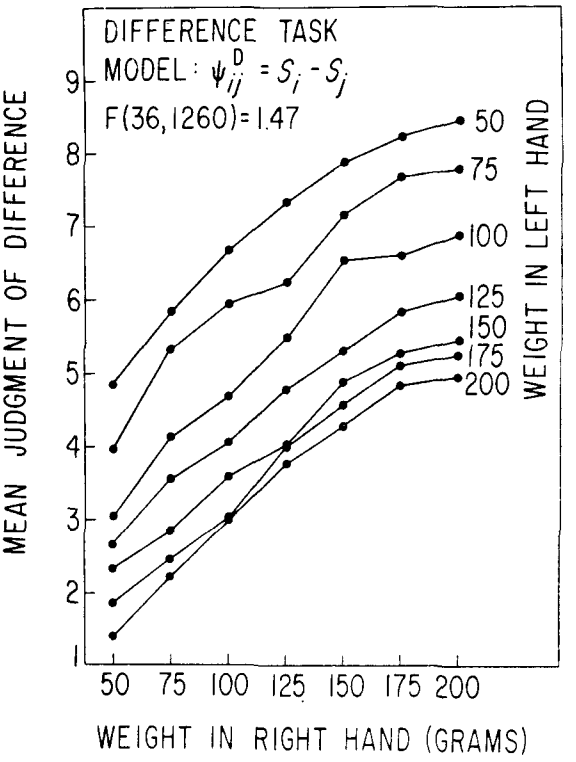

Fig. 2. Mean judgments of difference in heaviness between right- and left-hand weights, as a function of the right-hand weight. Separate curves represent different left-hand weights.

Subjects

The Ss were 24 UCLA undergraduates, fulfilling a requirement in introductory psychology.

Twelve naive geophysics students, who were acquaintances of E, were run in a preliminary study, performing the difference and ratio tasks only. They were run in a similar fashion as the 24 Ss in the main experiment. The category scale was reversed for the difference task; these ratings were subtracted from 10 in order to make the scale agree with that of the main experiment. Because the data for these 12 Ss agreed in every respect with those of the main experiment, their data were combined with the 24 Ss in the main study.

RESULTS

Difference Task

Figure 2 plots mean judgments of differences, averaged across Ss, as a function of the weight in the right hand, with a separate curve for each level of left-hand weight. The vertical separations between the curves represent the effects of left-hand weight; the slopes represent the effects of right-hand weight. Thus, positive slopes simply indicate that judgments of the difference in heaviness between the right hand and left hand increases as level of weight in the right hand increases.

The subtractive model predicts that the curves should be parallel, since the differences between the curves are a

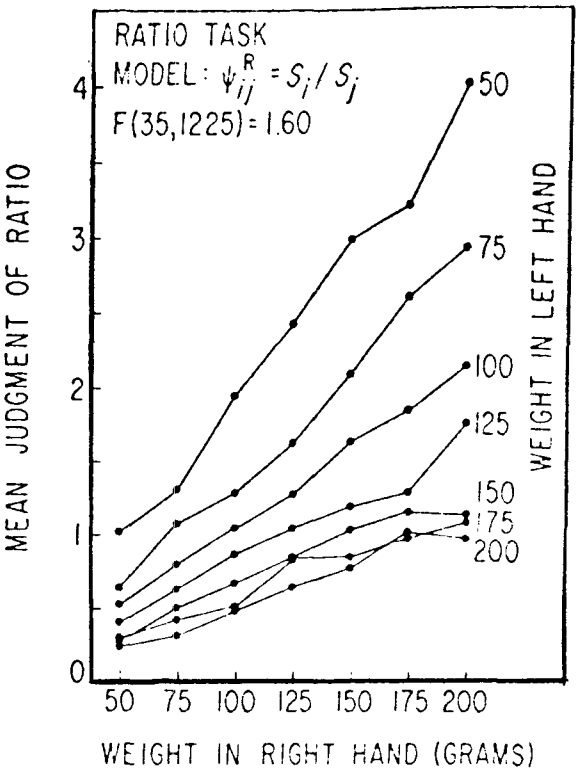

Fig. 3. Mean judgments of ratios as a function of level of right-hand weight.

function of left-hand weight only. As can be seen in the figure, the curves are very nearly parallel. The mean absolute discrepancy from the subtractive model is only 1/10 of one category. The analysis of variance test for interaction (nonparallelism) was small, though of borderline statistical significance, F(36,1260) = 1.47. Considering the power of the test and the lack of any systematic discrepancy from parallelism in Fig. 2, the results support the subtractive model.

It thus appears that the results are consistent with the assumptions that Ss compute differences when instructed to do so, and that the judgment function, J, is linear for category ratings of differences.

Ratio Task

The mean ratio estimations divided by 100 are shown in Fig. 3, plotted in the same way as in Fig. 2. Since the ratio model of Eq. 2 is a multiplicative model, the graphic prediction is one of a family of diverging curves which intersect at a common point (Anderson, 1970). Statistically, the interaction should be significant and located entirely in the bilinear component.

The interaction is, in fact, divergent and highly significant, F(36,1260) = 22.53, with 82% of the variance in the bilinear component. As can be seen in the figure, the data follow the graphic prediction of a bilinear fan of curves. The residual from bilinearity was small, though of borderline significance, F(35,1225) = 1.60, indicating that the ratio model gives a good approximation to magnitude estimations of "ratios."[1]

Averaging Task

Figure 4 plots mean judgments of averages as a function of level of left-hand weight, with a separate curve for each level of right-hand weight. Judgments of average heaviness increase with increases of weight in both right and left hands.

The constant-weight averaging model, as a special case of an additive model, predicts parallel curves for Fig. 4. Instead, the figure shows that the curves converge toward the right, in violation of this prediction. A test of the interaction was statistically significant, F(36,828) = 2.49, with 54% of the interaction variance concentrated in the bilinear component. The test of the bilinear interaction is highly significant, F(1,23) = 14.94.

The convergent interaction is similar to that observed in a number of studies involving averaging of psychophysical stimuli, for loudness averaging (Parducci, Thaler, & Anderson, 1968), length averaging (Birnbaum, Parducci, & Gifford, 1971), and averaging of motor movements (Levin, Craft, & Norman, 1971).

The present results provide further evidence against a simple averaging model, since the effect of any weight appears to depend on the weight with which it is paired.[2] However, the possibility remains that the discrepancies reflect nonlinearity in the J function for the rating response. Thus, it is possible that the constant-weight averaging model is an appropriate

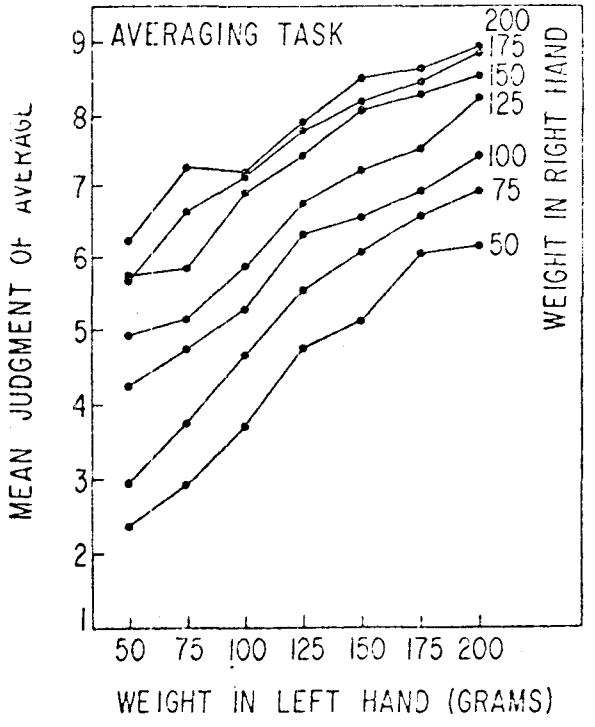

Fig. 4. Mean judgments of averages as a function of left-hand weight, with a separate curve for each level of right-hand weight.

description of the integration function, I, but that the overt responses must be transformed to eliminate response bias. The propriety of rescaling the data can be examined with respect to the scale convergence criterion.

Scale Convergence Criterion Applied

Briefly, the above findings suggest that when instructions specify "differences," the subtractive model fits; when instructions specify "ratios," the ratio model fits; however, when the instructions are to judge "averages," the simple averaging model does not fit. According to a simplistic view of functional measurement, the fit of the models for the first two tasks would simultaneously validate the models and the two response scales. The interpretation of the averaging task would be uncertain--either the response scale or the simple averaging model (or both) are not valid.

The next section shows how the scale convergence assumption can help remove some of the uncertainty for the averaging task: rescaling of the averaging data would be considered appropriate if, after rescaling, scale values derived from the model agreed with those from the other tasks. The section after next shows how comparison of scale values for the difference and ratio tasks challenges the simplistic view of functional measurement.

Averaging vs Difference Task Scales

If we assume the validity of the subtractive model for the difference task, then the parallelism of the raw data in Fig. 2 supports the conclusion that the J function for category ratings of differences in this situation is linear. Hence, the scale values derived from the model (marginal means) form an interval scale of the stimuli. Thus, each curve in Fig. 2 traces the shape of the psychophysical function, H, for the subtractive model. The curves are negatively accelerated and could be roughly approximated by either a logarithmic relation or a power function with a low exponent (.18). This exponent differs drastically from results obtained with magnitude estimations (Stevens & Galanter, 1957), which implied exponents greater than one. However, exponents less than one (approximately .6) have been obtained by investigators using additive and subtractive models (e.g., Rule, Curtis, & Markley, 1970; Anderson, 1972).

Since the averaging model did not fit the data, it is more difficult to find scale values. However, when the weights in the two hands are of equal value, almost any model will imply that the impression of heaviness equals the scale value. Judgments of equal pairs, R^A_{ii}, are therefore $J(s_i)$, where J is the judgment function for category ratings of averages. The scale convergence criterion provides a method for finding J, and for *appropriate* rescaling, if the data so dictate.

Figure 5A plots R^A_{ii} as a function of the subtractive

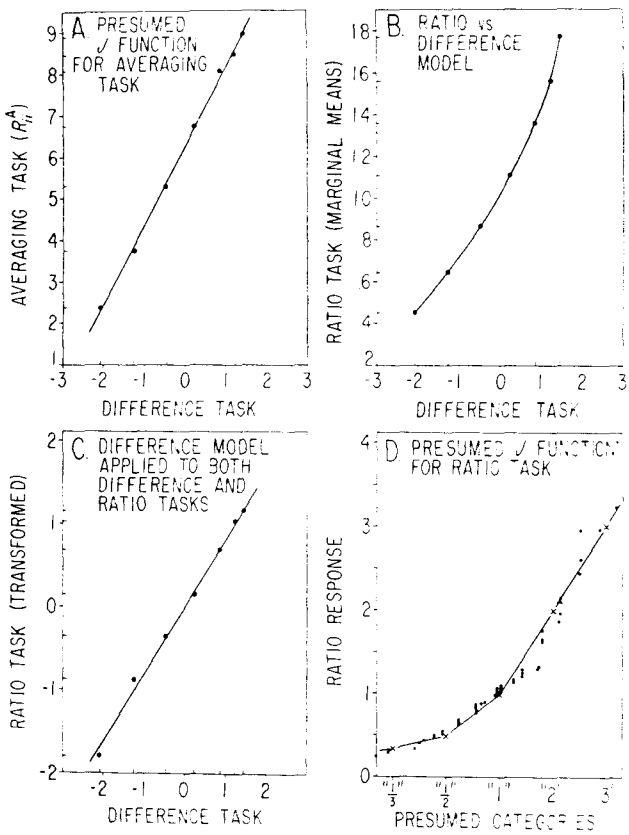

Fig. 5. (A) Mean judgments of equal-weight pairs for the averaging task as a function of the difference model scale values. (B) Scale values for the ratio model as a function of the difference model scale values. (C) Relationship between scale values derived from application of the subtractive model to the data for both the difference and ratio tasks. To apply the subtractive model to the ratio task data, the responses were transformed to additivity, using MONANOVA. (D) Presumed J function for the ratio task. The solid curve depicts the relationship between the categories representing "ratios" and the numerical response required of the S. The points show the transformation estimated by MONANOVA that makes the ratio data fit the subtractive model.

model scale values, derived from the difference task data. The abscissa values are assumed to represent s and the ordinate values are assumed to represent J(s). Therefore, the curve in Fig. 5A represents an almost perfectly linear judgment function for the averaging task. Thus, these data do not dictate transformation. Indeed, if the validity of the subtractive model is assumed, the linear relationship in Fig. 5A validates the rating scale for the averaging task. Also, this result then implies that the nonparallelism in Fig. 4 is "real," not due to a response bias.

A subsidiary analysis checked the possibility that a mild monotonic rescaling might make the averaging data in Fig. 4 parallel, while retaining scale values compatible with the other tasks. The mean averaging data were transformed to parallelism using MONANOVA, a computer program based upon Kruskal's (1965) monotone rescaling procedure (Kruskal & Carmone,

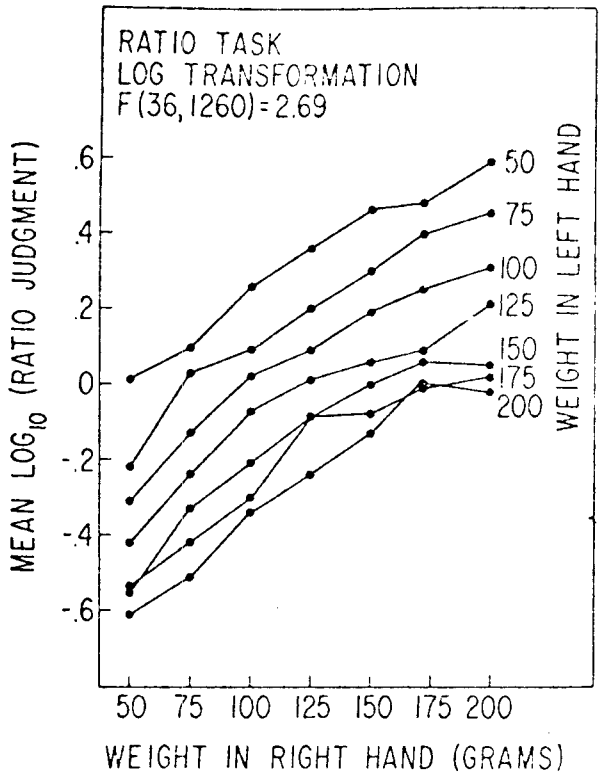

Fig. 6. Mean logarithm of ratio estimations for the ratio task, plotted as in Figs. 2 and 3.

1969). Following a positively accelerated transformation, the scale values derived in accordance with the constant-weight averaging model were nonlinearly related to both the difference and ratio task scales. Response scaling would therefore be considered inappropriate if the convergence criterion is accepted. Thus, the interaction is "real," and one must seek alternative integration functions, such as either the range model (Birnbaum et al, 1971) or differentially weighted averaging models (Anderson, 1971).

Ratio vs Difference Task Scales

If the validity of the ratio model and the linearity of J for the ratio task are assumed, then the marginal means are estimates of the scale values for the right-hand weight and of reciprocals of the scale values for the left-hand weight. The curves in Fig. 3 describe the shape of the psychophysical function for the ratio model, indicating a near-linear relationship with a slight positive acceleration.

Figure 5B plots the scale values derived from the ratio model as a function of the difference task scale values. The function is markedly nonlinear, showing a positively accelerated relationship between the two sets of scale values. This exponential relationship was found for all but two individual Ss. Although both models agree with the data for their tasks, their respective scales do not agree.

Scale convergence can provide a basis for response transformation, even though both models fit the raw data. The problem is to find representations of *both* sets of data, so that the two sets yield compatible scale values. The answer to this problem, explained by analyses below, is that the *same I* function applies to both tasks.

The ratio task data were transformed to fit the subtractive model by means of MONANOVA. Figure 5C plots the scale values obtained from this analysis as a function of the subtractive model scale values derived from the difference task data. This plot is almost perfectly linear, consistent with the interpretation that Ss are performing the *same* integration function for both tasks, regardless of instructions. The inverse of the monotone transformation that eliminated the interaction is shown by the solid points in Fig. 5D. If "ratios" are really computed by subtraction, then the points in Fig. 5D depict the J function for magnitude estimations of "ratios." The J function estimated by this procedure is nearly exponential.

The ratio task data were transformed by a logarithmic function, which would counteract the effect of the presumed exponential J function. The mean log ratio response is shown in Fig. 6. Although the interaction is statistically significant, $F(36,1260) = 2.69$, the curves appear roughly parallel and do not show any systematic deviations from parallelism. The 49 points in Fig. 6 were a nearly perfect linear function of the mean ratings in Fig. 2, indicating that the responses for the two tasks were indeed nearly exponentially related. In addition, the marginal mean log ratios were linearly related to the subtractive model scale values for all but two individual Ss.

These findings eliminate the possibility that both difference and ratio models apply with a common scale. If $R^R = (s_i/s_j)^\alpha$, then $\log R^R = \alpha[\log s_i - \log s_j]$. Hence, the marginal mean logs would be a logarithmic, rather than linear, function of subtractive model scale values. The same argument applies to the MONANOVA results shown in Figs. 5C and 5D. If the ratio model were correct and J was a power function, the points in Fig. 5C would be logarithmic rather than linear. MONANOVA gives the same scale values to both sets of data, because the responses for the two tasks, R^R and R^D, are monotonically (exponentially) related.

The findings that the raw data fit the simple ratio and subtractive models and that scales agree only when both sets of data are fit to the same model suggest that one of the simple models applies to both tasks.

It should be emphasized that the present data suggest that Ss perform the same I function for "differences" and "ratios," but these data *do not* permit specification of what the I function is. This point deserves elaboration. The data are compatible with either of the following theories: (1) $R^D = a(s_i - s_j) + b$, and $R^R = \exp(s_i - s_j)$; *or* (2) $R^D = a\log(s_i/s_j) + b$, and $R^R = (s_i/s_j)^\alpha$. The data are *not* compatible with the following:

$R^D = J[s_i - s_j]$ *and* $R^R = J^*[s_i/s_j]$, where J and J* are any arbitrary monotonic functions. This follows because $R^R = \exp(R^D)$ for these data. For these two integration functions to apply with the same H function, the responses must not be monotonically related (Krantz et al, 1971; Krantz, 1972).

If the scale convergence criterion is accepted, we must face the indeterminacy that Ss integrate information in the same way for "differences" and "ratios." One approach to the problem would be to propose testable theories of the J function. If we assume the validity of the ratio model for both tasks, we must conclude that the category ratings induce an *exactly* logarithmic J function. However, no testable theory for this relation seems evident.

On the other hand, to conclude that the subtractive model applies to both tasks would imply that the J function for magnitude estimation would have to be exponential. The following speculation offers a testable and potentially useful explanation of magnitude estimation: Suppose Ss treat magnitude estimation "ratios" as equal-interval categories. When given a scale of "ratios," they consider the differences between adjacent "ratios" to be equal, as though they were categories. Hence, Ss treat the psychological difference between "four times as heavy" and "twice as heavy" as being equal to the difference between "one-half as heavy" and "one-fourth as heavy." Magnitude estimation instructions typically include several examples, such as "if it seems twice as heavy, say '200,' if it seems four times as heavy, say '400.'" If the examples defined the scale for ratios of 1/4, 1/2, 1, 2, and 4, then the transformation would be perfectly exponential. Since these examples are frequently used, it should be no surprise that magnitude estimation scales are often related by an exponential transformation to category scales. Hence, the particular J function for a given magnitude estimation experiment may depend upon the examples used to define the scale as well as a variety of other factors (Poulton, 1968)

DISCUSSION

These data suggest that when Ss are instructed to rate "differences," the ratings satisfy the parallelism dictated by the subtractive model (Fig. 2). When instructions specify "ratios," the data conform to the bilinear prediction of the ratio model (Fig. 3). However, the scale values derived from the two sets of data agree only when both sets of data are fit to the *same* model. These results are in startling agreement with Torgerson's (1961) conjecture that Ss do not distinguish "differences" from "ratios."

Functional Measurement

According to a simplistic view of functional measurement, the fit of the model simultaneously establishes the validity of the model and the response scale. However, if only the difference task or only the ratio task had been studied, the investigator using this simplistic approach would have concluded that both the model and response scale were validated. By studying both tasks in the same experiment, by assuming that the psychophysical function is task invariant, and by limiting the discussion to the simple subtractive or ratio models, we conclude that at least one model and one response scale are not valid. The exponential relationship between the two response procedures allows both models to fit when only one integration function appears to operate.

A more intricate view of measurement and model testing is required. A further constraint is required to specify all of the functions of functional measurement (Fig. 1B) and remove the lingering indeterminacy. It is not possible to simultaneously validate *both* the model and the response scale in a single two-factor experiment. These problems do not seem insurmountable, however, and three lines of attack seem to offer great promise: First, a coherent system of models and response scales over a wide variety of situations, together with the concept of simplicity, may suffice (Anderson, 1972; Birnbaum, 1972). Second, the scale-free approach as applied in Birnbaum (1972, Experiment IV) may provide additional insight into the integration process. Third, theorization about the judgment function can lead to testable consequences, which, if verified, would predict and explain the actual responses obtained in a given experiment. This more intricate view is very much in the spirit of functional measurement, with its concern for explaining overt responses rather than merely ordinal relationships, but it requires more ordinal constraints.

Magnitude Estimation May Be Valid

Anderson (1972), based on a variety of lines of evidence (cf. Weiss, 1972; Curtis, Attneave, & Harrington, 1968; Curtis & Fox, 1969; Curtis, 1970; Rule, Curtis, & Markley, 1970), has recently argued that magnitude estimation "must be biased and invalid." It must be emphasized that this conclusion follows only from the assumption that the integration processes studied in the above references were additive.

The arbitrariness of this conclusion is illustrated by Weiss's (1972) direct comparison of magnitude estimation with graphic rating for an averaging task. When graphic ratings were employed, the data fit the constant-weight averaging model; however, the magnitude estimations were quantitatively consistent with the geometric averaging model.[3] Only the assumption of one model or the other would lead to the conclusion that one response procedure was "valid" and the other "biased."

If we assume the validity of the constant-weight averaging model for the Weiss (1972) study, we must conclude that magnitude estimation induces an

exponential transformation. Similarly, if the validity of the subtractive model is assumed for the present ratio task, then an exponential J function would be required to explain the fit of the ratio model (Fig. 3).

However, if we assumed a geometric averaging model for the Weiss (1972) study and a ratio model for the present ratio task, we might conclude that magnitude estimation is a valid procedure. Thus, it may be possible to develop a coherent system based on multiplicative models and magnitude estimation. To call one response measure the "true measure of sensation" seems premature at present.

CONCLUDING COMMENTS

The simplest interpretation appears to be that the subtractive model is the appropriate representation for both the difference and ratio tasks, that magnitude estimation induced an exponential judgment function, and that the interaction for the averaging task should be attributed to a nonadditive integration function rather than to a bias in the judgment function. These conclusions follow from the scale convergence criterion and the assumption of the subtractive model for the difference task. (1) The parallelism of the curves in Fig. 2 supports the linearity of J for ratings of differences. (2) That ratings of averages of equal weights are linearly related to subtractive model scale values (Fig. 5A) validates the rating scale for the averaging task. (3) The findings that scale values for the difference and ratio tasks are exponentially related (Fig. 5B), and that scale values for the two sets of data agree when both are fit to the same model (Fig. 5C) suggest that Ss use the same I function for "differences" and "ratios." (4) The fit of the ratio model (Figs. 3 and 6), together with the assumption that I is subtractive, implies an exponential J function for the ratio task.

Scale convergence can thus prevent blaming the response scale when a model fails (averaging task) and can dictate rescaling even when the model fits (difference and ratio tasks). It could also dictate appropriate rescaling to rectify nonlinearity in the response scale. Scale convergence seems a reasonable requirement and a potentially useful concept for the study of psychological laws. However, one must continue to question even the most reasonable assumptions. This invariance assumption can be considered a provisional criterion that is also a potentially testable empirical proposition.

Further research is needed to examine the generality of these results for different stimulus continua. For certain continua, such as visual length, it seems intuitively plausible that Ss could judge both differences and ratios. The study of a variety of tasks involving different algebraic models seems a promising approach to understanding information processing and judgment.

REFERENCES

Anderson, N. H. On the quantification of Miller's conflict theory. Psychological Review, 1962, 69, 400-414.

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

Anderson, N. H. Information integration and attitude change. Psychological Review, 1971, 78, 171-206.

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

Birnbaum, M. H. The nonadditivity of impressions. Unpublished doctoral dissertation, University of California, Los Angeles, 1972.

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

Bogartz, R. S., & Wackwitz, J. H. Polynomial response rescaling and functional measurement. Journal of Mathematical Psychology, 1971, 8, 418-443.

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

Curtis, D. W. Magnitude estimations and category judgments of brightness and brightness intervals: A two-stage interpretation. Journal of Experimental Psychology, 1970, 83, 201-208.

Curtis, D. W., Attneave, F., & Harrington, T. L. A test of a two-stage model for magnitude estimation. Perception & Psychophysics, 1968, 3, 25-31.

Curtis, D. W., & Fox, B. E. Direct quantitative judgments of sums and a two-stage model for psychophysical judgments. Perception & Psychophysics, 1969, 5, 89-93.

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

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

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

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

Kruskal, J. B. Analysis of factorial experiments by estimating monotone transformations of the data. Journal of the Royal Statistical Society (B), 1965, 27, 251-263.

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

Levin, I. P., Craft, J. L., & Norman, K. L. Averaging of motor movements: Tests of an averaging model. Journal of Experimental Psychology, 1971, 91, 287-294.

Parducci, A., Thaler, H., & Anderson, N. H. Stimulus averaging and the context for judgment. Perception & Psychophysics, 1968, 3, 145-150.

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

Rule, J. S., Curtis, D. W., & Markley, R. P. Input and output transformations from magnitude estimation. Journal of Experimental Psychology, 1970, 86, 343-349.

Seward, J. P. The constancy of the I-V: A critique of intervening variables. Psychological Review, 1955, 62, 155-168.

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

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

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

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

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

Treisman, M. Sensory scaling and the psychophysical law.

Quarterly Journal of Experimental Psychology, 1964, 16, 11-22.

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

Zinnes, J. L. Scaling. Annual Review of Psychology, 1969, 20, 447-478.

NOTES

1. Two other analyses supported the fit of the ratio model. (1) The first test multiplied each ij^{th} and ji^{th} entry of the response matrix for each S. From Eqs. 2 and 4, these products can be written

$$X_{ij} = [a(s_i/s_j) + b] \cdot [a(s_j/s_i) + b]$$

$$= a^2 + ab(s_i/s_j) + ab(s_j/s_i) + b^2,$$

which depends upon s_i and s_j. However, if b = 0, then $X_{ij} = a^2$, indicating that the products should be constant for all cells in the design. The mean of these products was 1.02, and they did not appear to show any systematic pattern of deviations from this value. (2) If the intercept (b) is zero, the product of each row marginal mean, cs_i and corresponding column marginal mean, $d(1/s_i)$, should be equal to a constant, cd, providing the scale value of each stimulus is independent of which hand it is presented in. However, if b were nonzero, then these products would be a monotonic function of the level of weight. These products were taken for each S and level of weight, and they were not significantly different from one another, $F(6,210) = 1.54$, nor were they a monotonic function of the level of weight; hence, the linear trend was also nonsignificant, $F < 1$.

2. The range model (Birnbaum et al, 1971) was fit to the averaging task data. The simple range model, $\Psi_{ij} = .5(s_i + s_j) + \omega \mid s_i - s_j \mid$, can be rewritten as a configural-weight model, $\Psi_{ij} = (.5 + \omega)s_i + (.5 - \omega)s_j$, when $s_i > s_j$. Estimating one additional parameter (ω), the range model provided a highly significant improvement over the constant-weight averaging model, $F(1,35) = 29.27$. An averaging model with differential weights fit nearly as well as the range model, although it used a greater number of parameters.

3. It is possible that the change in instructions in the Weiss (1973) study may have induced the Ss to use a different integration rule. However, a subsequent analysis of these data based on the scale convergence criterion rules out this possibility. The psychophysical scales derived from the two sets of data were more nearly linearly related when both sets were transformed to fit the same I function.

(Received for publication August 21, 1972; final revision accepted July 27, 1973.)