Cross-task validation of functional measurement*

NORMAN H. ANDERSON

University of California, San Diego, La Jolla, California 92037

Ss judged average heaviness of two unseen lifted weights on a 20-step rating scale. They also rated heaviness of a single visible object of variable size and weight. The data for each task obeyed the parallelism prediction of the linear integration model. This within-task consistency validates jointly the model and the response scale. Functional measurement procedure yielded independent scales of subjective heaviness from each task. These were linearly related, a criterion of external consistency. The plot of subjective heaviness against objective weight was slightly negatively accelerated. These results disagree with those that have been obtained with magnitude estimation. Since the rating data satisfy both internal and external consistency, they constitute the true measure of sensation. Magnitude estimation, in contrast, must be biased and invalid.

An important consequence of developments in functional measurement theory is the cumulation of evidence in support of ordinary rating responses (Anderson, 1970a, 1972). This approach relies on simple algebraic models for stimulus integration. Averaging, adding, subtracting, and multiplying models have been successfully used in various applications with rating responses. If the rating response was not on an interval scale, that would invalidate a correct model. Success of the model thus constitutes a validation, not only of the model itself, but also of the rating response.

But rating scales have been repeatedly condemned by S. S. Stevens (e.g., 1957, 1971). Stevens has claimed that ratings are biased, whereas the method of magnitude estimation gives a correct response measure. Stevens has certainly shown that these two response measures are nonlinearly related for certain dimensions, such as heaviness. Both cannot be correct, therefore, and the choice between them has considerable importance. If Stevens were correct, then much of the work on stimulus integration would need to be reconsidered. Accordingly, this issue needs careful consideration.

Stevens, as is well known, fails to provide a validation criterion for magnitude estimation. Fitting a power function to the data does not constitute a criterion, if only because Marks (1968) has shown that power functions will fit rating data also.

*This work was supported by National Science Foundation Grant GB-21028 and by grants from the National Institute of Mental Health to the Center for Human Information Processing, University of California, San Diego. I wish to thank Allen Parducci and David Weiss for their comments. I also wish to thank Cliff Butzin for his help in running these experiments.

Crossmodal matching does not constitute a criterion, if only because it can occur with a nonlinear response scale (see below; also Anderson, 1970a, p. 166). Stevens appears to recognize this situation when he refers to scale validity as "a matter of opinion" (Stevens, 1971, p. 448).

In functional measurement, scale validity is not "a matter of opinion." The algebraic integration model provides a criterion of internal consistency. This constitutes a validation base for the response measure. Ratings and magnitude estimates have equal opportunity. The success of the ratings means that they are an interval response measure. Since magnitude estimates are nonlinearly related to the ratings, it follows that magnitude estimation is biased. Magnitude estimation does not even yield an interval scale, much less a ratio scale. A striking empirical demonstration is given by Weiss (1972).

This report takes the next step. It tests the external consistency of the functional stimulus scales. Two integration tasks are used with lifted weights: weight averaging and the size-weight illusion. The same stimulus scale should be obtained in both tasks. Since the two tasks involve rather different integration processes, this cross-task comparison has special interest.

Both these integration tasks have been studied before. In two experiments (Anderson, 1967; Anderson & Jacobson, 1968), the S lifted several weights in succession and rated average heaviness of the sequence of weights. The data supported the averaging model, though not without some reservation about a possible floor-ceiling effect in the second experiment. A third experiment (Anderson, 1970b) gave good support to a linear model for

heaviness ratings of objects varying in weight and size. This report studies both tasks together and uses more levels of the weight dimension to get a more extensive stimulus scale.

METHOD

Weight Averaging

In this task, the S lifted two weights in succession and judged their average heaviness. A 53 x 91 cm panel hid the weights and the E. On the S's side of the panel, two handles led by strings over pulleys to trays on which the weights were placed. The S was instructed to pull each handle down briefly once, in right-left order, using his right hand only. Responses were on a 1-20 rating scale.

Each weight could have one of five values: 200, 300, 400, 500, and 600 g, including the weight of the tray. This 5 by 5 design yielded 25 weight pairs. In addition, there were two end anchor pairs, in which the two weights were both 100 g or both 700 g. These anchor pairs were presented twice during the initial instructions and twice before each replication of the 25 pairs. Each time an anchor was presented, the S was told to call it 1 or 20, respectively, for the light and heavy anchors.

Four replications of the 25 pairs were given, in separate randomized order separated by short rests. The S was told that the first replication was practice, and only the data from the last three replications are reported here.

Size-Weight Illusion

In this task, the S judged the heaviness of a single visible weight. Each weight was a gray plastic cylinder, 6 cm in diam, with a lifting loop attached by wires to three points trisecting the circumference of the lid. One handle of the above apparatus was used, but with the S shifted 90 deg to make the weight visible. The tray was removed, and the E attached the weights to a hook on the end of the string. The S then pulled down briefly once on the handle as in the first task.

The weights were constructed according to a 5 by 5, Size by Weight design. The cylinder heights ranged from 3.3 to 15.3 cm in 3-cm steps. The cylinders were weighted with lead and clay to produce weights of 200, 300, 400, 500, and 600 g, as in the first task. The two end anchors were a 100-g weight of height 3.3 cm and a 1,000-g weight of height 6.3 cm, to be called 1 and 20, respectively. Four replications of the set of stimuli were given, using the same procedure as in the first task.

The Ss were 20 right-handed students from introductory psychology. They served in both tasks,

Fig. 1. Mean rated heaviness for two judgment tasks. Weight averaging in left panel, and size-weight illusion in right panel.

in the above order, in a single session of about 70 min. No information about the size-weight illusion was given.

RESULTS

Integration Models

Averaging model. The left panel of Fig. 1 plots the mean response from the averaging task. The averaging model implies that the five curves should be parallel. Since this parallelism prediction is well supported, it is concluded that the *S* does indeed average the subjective heaviness values of the two weights. Strictly speaking, any linear integration model with constant weight parameters will predict parallelism. To show that the *S* averages, rather than adds, requires collateral data (see Anderson, 1972).

The simplicity of the parallelism test deserves emphasis. To test the model requires only a plot of the raw data. No parameter estimation is required. Even the spacing on the horizontal axis is arbitrary.

A rigorous test of fit is also available. The parallelism prediction is equivalent to a prediction of zero interaction in the analysis of variance test. That interaction was not significant, $F(16,304) = 1.04$. The *F* ratios for first and second weights were nearly the same, indicating that the two weights played a nearly equal role in the overall response to the pair.

Size-weight illusion. The right panel of Fig. 1 shows the mean response from the second task. The linear integration model again yields a parallelism prediction. Except for one deviant point, for the lightest, smallest

weight, these data also follow the parallelism prediction. Here again, it deserves emphasis that a simple plot of the raw data provides a direct, transparent test of the model.

The deviant point in the right panel of Fig. 1 may have resulted from the use of a light end anchor of the same 3-cm height as the smallest weight. That could have produced a few excess "1" responses in direct response to the size cue. Regardless of its origin, this deviant point was enough to produce a significant interaction, $F(16,304) = 2.05$. With that point omitted, the interaction is not significant, $F(15,304) = 1.10$. The sensitivity of the analysis of variance test to this one-point discrepancy is noteworthy.

Functional Stimulus Scales

Functional measurement theory implies that the vertical spacing of the curves in each panel of Fig. 1 constitutes an interval scale of subjective heaviness. This can be illustrated with the model for the size-weight illusion:

$$R_{ij} = w_1 s_i + w_2 s_j^*, \quad (1)$$

where R_{ij} is the response to Weight *i* of Size *j*, s_i is the scale value (subjective heaviness) of Weight *i*, s_j^* is the scale value of Size *j*, and w_1 and w_2 are mathematical weight parameters that reflect the relative importance of the weight and size cues on the response.

The difference between the two top curves in the right panel of Fig. 1 is thus

$$R_{600,j} - R_{500,j} = w_1(s_{600} - s_{500}). \quad (2)$$

The parallelism prediction, it may be noted, follows directly from Eq. 2, because the expression on the right is independent of the size cue.

Of more immediate relevance is that any two curves yield a relation similar to Eq. 2. The difference between any two curves thus is proportional to the difference between the scale values of the two corresponding weights. In other words, the vertical spacing of the curves constitutes an interval scale of subjective heaviness.

Each judgment task yields an independent functional scale of heaviness. They should be linearly related. The test is shown in the left panel of Fig. 2, which plots the heaviness scale from the size-weight task on the vertical against the heaviness scale from the weight-averaging task on the horizontal. Since the relation is linear, this graph provides a cross-task validation of the stimulus scales and of the response scales.

For comparative purposes, the right panel of Fig. 2 plots the functional scale for the second weight against the functional scale for the first weight in the averaging task. These too are linearly related. Their average was used in the panel on the left.

The Psychophysical Law

In the functional measurement approach, stimulus scales are derivative from the psychological law, i.e., from the integration model. The psychophysical law is merely the relation between these subjective stimulus values and the physical stimulus values.

These graphs are shown in Fig. 3 for both judgment tasks. Both curves show a slight negative acceleration, in agreement with previous work in this program (Anderson, 1970b; 1971, p. 60). Since negative acceleration would imply an exponent less than 1 in a power law, these results disagree with those from magnitude estimation, which yields an exponent around 1.45 (Stevens & Galanter, 1957).

A functional scale of the size cue is also available. It can be visualized as the mean of the five curves in the right panel of Fig. 1. This scale shows a negative acceleration, with successive height increments having diminishing effect. That is consistent with previous findings of linearity as a function of the cube root of the volume of cubes (Anderson, 1970b). Cubes are probably preferable to cylinders, since they give a larger effect.

The scale for the size cue has special theoretical interest. It is a heaviness

value, derivative from the size cue by the presumed operation of a learned expectancy. This is the functional value of the size cue, i.e., the value that is immediately operative in the stimulus integration. It would be difficult to obtain in any other way. Present methods apply to objects of any size and appearance.

DISCUSSION

Stimulus Integration

Weight-averaging and size-weight judgments both follow simple integration models. In each case, the S has two cues; and in each case the cues combine in a simple linear manner. These results agree with previous reports on weight averaging (Anderson, 1967; Anderson & Jacobson, 1968) and on the size-weight illusion (Anderson, 1970b).

The subjective heaviness scale was found to be the same in both judgment tasks. The importance of this cross-task validation is increased by the difference in the integration processes in the two tasks. Of itself, the averaging model may seem trivial, but it has a useful function (Weiss & Anderson, 1969, p. 61; 1972). In the present report, the success of the model for weight-averaging buttresses the linear model for the size-weight illusion. The popular interpretation, that Ss really judge density, yields a nonlinear model (Anderson, 1970b).

No less important is the concomitant validation of the rating response. The parallelism prediction verified in Fig. 1 requires an interval

response measure. If the response was not on an interval scale, then the curves would not be parallel even though the models were correct.

Magnitude Estimation

Present results imply that magnitude estimation is biased and nonlinear. It does not even provide an interval scale, much less a ratio scale. As Stevens and Galanter (1957) have shown, and as Stevens (1971) has repeatedly emphasized, magnitude estimation is nonlinearly related to ordinary ratings, for lifted weight and for certain other dimensions as well. Nonlinear transformation of the rating responses of Fig. 1 would in general destroy the parallelism. If the integration models are correct, then the magnitude estimation procedure as used by Stevens is invalid.

Two other reports bear on this point. J. C. Stevens and Rubin (1970) studied the size-weight illusion using magnitude estimation. In their Table 1, the curves for fixed gram weight diverge as size decreases. A nonlinear response scale would produce nonparallelism, and their obtained divergence interaction is exactly what would be expected if magnitude estimation has the bias discussed below.

In addition, Weiss (1972) has made a direct comparison between magnitude estimation and rating procedures. Weiss studied averaging of grayness under conditions similar to those used here. His rating data obey the parallelism prediction, and thus provide joint support for the averaging

Fig. 3. The psychophysical law for heaviness. Subjective heaviness is plotted against the objective weight scale for the two judgment tasks.

model and for the rating scale. However, the magnitude estimation data exhibit a strong divergence interaction.

The nature of the bias in magnitude estimation is easily seen and deserves closer consideration. Various writers have suggested that the number continuum obeys a law of "diminishing returns," so that the subjective difference between 10 and 20, say, is larger than that between 110 and 120. Such a relation, it may be noted, would be consistent with the inverse correlation of everyday life between value and numerosity. Nonlinearity in the number continuum could thus represent experiential factors, analogous to the size cue to heaviness.

Such a nonlinear bias would account directly for the cited divergence interaction in the magnitude estimation data. To express a subjective difference corresponding to that between 10 and 20, the S would have to use two numbers farther apart than 110 and 120. A linear model would then yield nonparallel curves that diverged with increasing intensity.

A more penetrating analysis of the bias in magnitude estimation can be obtained by following out a line of argument introduced by Attneave (1962). To a rough approximation, the relation between subjective and objective number, n and N , could be represented by a power function with an exponent less than 1:

$$n = N^a, a < 1. \quad (3)$$

Stevens's entire conceptual edifice, as

Fig. 2. Comparison of functional stimulus scales. Left panel plots heaviness scale from size-weight illusion on the vertical as a function of the heaviness scale from the weight-averaging task on the horizontal. Right panel shows comparative data for the two heaviness scales obtainable from the weight-averaging task.

is well known, rests on the assumption that $a = 1$.

Two implications flow from Eq. 3. First, suppose that subjective stimulus value, s , is a power function of the physical stimulus value, S , with exponent, b . Since N is the observable in magnitude estimation, it follows that

$$N = Sb^{1/a}. \quad (4)$$

Equation 4 means that b/a is the exponent that will be obtained from magnitude estimates. Since $a < 1$ by assumption, these observed exponents, b/a , would be greater than the actual or correct exponents, b .

The present results show a near-linear relation between subjective and objective weight. Thus, the value of b would be near 1 for lifted weight. The above argument implies, therefore, that the power function for lifted weight should have an exponent greater than 1 if magnitude estimation is used. The value of 1.45 used by Stevens would correspond to an a value around .7, a serious nonlinear bias.

The other implication of Eq. 4 is worth noting. Stevens has attempted to supply a validation criterion for magnitude estimation by crossmodal matching. In addition to judging each of two stimulus dimensions relative to the number continuum, they may be judged relative to each other. Any one exponent should then be a ratio of the other two. But that will be true regardless of the value of a in Eq. 3. Crossmodal matching does not validate magnitude estimation.

The foregoing argument needs qualification in one important respect, since it would imply that magnitude estimation was necessarily biased and would always disagree with ordinary ratings. For certain physical stimulus dimensions, however, ordinary ratings and magnitude estimations appear to be linearly related (Stevens & Galanter, 1957). It is not appropriate, therefore, to treat numbers as essentially another physical stimulus dimension, nor is it possible to speak of a general bias or nonlinearity in the subjective use of numbers. A plausible interpretation is that S_s have a tendency to treat magnitude estimation essentially as a rating scale, and that this tendency is especially strong on certain stimulus dimensions. For example, Weiss and Anderson (1972) found that ordinary ratings and magnitude estimations of average angle were linearly related under the same conditions that gave a nonlinear relation for grayness averaging (Weiss, 1972). Inspection of the data showed that S_s under the magnitude estimation instructions simply used a

200-point rating scale centered around the standard value of 100. The presence of a partial tendency toward using magnitude estimation as an ordinary rating scale is also consistent with Poulton's (1968) model for the strong dependence of exponent on stimulus range. If magnitude estimation does indeed involve some sort of compromise between two number usage tendencies, it is unlikely to give clean data.

Other Weight Scales

Two other methods, fractionation and constant sum, have been used to determine weight scales by a number of investigators. Although these scales have been quite variable (see Baker & Dudek, 1955; Guilford, 1954; Stevens & Galanter, 1957, Fig. 6), they are typically convex upwards and have been taken to support the magnitude estimation procedure. Accordingly, these two methods require a brief discussion.

Fractionation. All applications of the fractionation method to lifted weights have assumed that the S employs that fraction, typically one-half, that is specified in the instructions. Unfortunately, that assumption is not generally testable in the ordinary experimental procedure (e.g., Garner, 1954; Torgerson, 1958, p. 113; see also Anderson, 1970a, pp. 160-162). If it is not correct, the fractionation scale will be biased, and the bias can be considerable. The scales that have been constructed by the fractionation method do not, therefore, bear on the problem at hand.

To illustrate this problem, suppose that subjective value is a power function of objective value and that the response obeys the fractionation model. Then the gram ratio between the weight judged half and the standard will be $w^{1/a}$, where w is the subjective fractionation constant and a is the exponent of the power function. This argument holds for any stimulus continuum (Anderson, 1970a, Eq. 15). Under the given assumptions, therefore, the observed physical fractionation ratio should be constant (see also Ekman, 1958, Eq. 3).

This constant ratio property can be seen to hold quite well in the data of Guilford and Dingman (1954, Table I) and Warren and Warren (1956). By the above argument, therefore, it is reasonable to conclude that the S_s were fractionating and that the psychophysical relation was a power function. On the other hand, the data of Harper and Stevens (1948) yield a nonconstant fractionation ratio that decreases irregularly from about .75 to .65. A power function would be inappropriate, unless perhaps a

threshold constant was allowed. In fact, Harper and Stevens fitted a log-log type of function.

But, though the fractionation method is testable under the power law assumption, it is inherently ambiguous about the subjective scale. The above theoretical expression, $w^{1/a}$, for the observable physical fractionation ratio, contains two unknowns. A scale can be obtained only if one unknown is assumed or known from other sources. The typical procedure has been to assume that the subjective fractionation ratio, w , is $\frac{1}{2}$. That allows a solution for a , the exponent of the power function, as in Guilford and Dingman (1954). Unfortunately, the fractionation method provides no way of testing whether w is $\frac{1}{2}$. The present data indicate that a is near 1 for weights, which implies that the observed gram fractionation ratio is approximately equal to the subjective fractionation ratio. The upward convexity in previous fractionation scales would seem to be a bias resulting from the assumption that $w = \frac{1}{2}$ when the true value is greater.

Warren and Warren, it should be noted, attempted to argue for a linear psychophysical relation using only fractionation data. They obtained a gram fractionation ratio of .64 when standard and comparison weights had equal volume, but this changed to .52 when the comparison weights were reduced to half the volume of the standard. This volume reduction was thought to correct for size-weight illusory effects, though that does not appear to be consistent with the present results on the psychophysical function for the size cue. More basically, their argument depends on the assumption that the subjective fractionation ratio is equal to the observed physical fractionation ratio, which is just what needs to be proved.

Constant sum. The constant-sum method and the ratio-estimation method as well have advantages over fractionation in terms of testability and scaling (Torgerson, 1958, p. 116). However, Guilford and Dingman (1954), Baker and Dudek (1955), Goude (1962), and Ottander (1967) all report certain inconsistencies between the data and the model. It is not clear whether the discrepancies in the model are serious, but such discrepancies do tend to vitiate the scales constructed from the model. This problem cannot be pursued here, but without further study the constant-sum scales have an uncertain value.

Weight-adding task. A brief note may be added on the weight-adding task used by Goude (1962). The S selected a weight whose heaviness

equaled the sum of the heavinesses of two other weights. Goude claimed that these data were consistent with the results from ratio estimation and yielded a power function with exponent 1.90, though his comparison is subject to certain reservations. Functional measurement procedure suggests a different approach to the sum production task.

The general model for sum production can be written

$$r_{ij} = w_1 s_i + w_2 s_j, \quad (5)$$

where r_{ij} is the subjective value of the response, s_i and s_j are the subjective values of the two stimuli to be summed, and w_1 and w_2 are mathematical weight parameters that can represent serial position effects. The model is written in subjective metrics since that is what the S uses in his processing. The observables, of course, are the corresponding physical stimuli. With a nonlinear psychophysical function, the observed response would require a monotone transformation to additivity (Anderson, 1962, 1970a; Weiss¹).

For lifted weights, however, the present finding of a near-linear psychophysical function implies that a response transformation may not be necessary. Equation 5 should be approximately correct in terms of the gram-response measures. Inspection of Goude's data (Table 4, p. 57) indicates reasonably good agreement with this analysis. For example, the first four points fall almost perfectly on a straight line.

Related Work

An interesting appearance of the simple parallelism test can be seen in Fig. 4 of Stevens (1971), which replots difference judgments of loudness from Beck and Shaw (1967). Although the data are somewhat variable and the stimuli are only approximately in a factorial design, the three sets of data points are roughly parallel. According to functional measurement, that supports both the subtracting model and the response scale. But, though magnitude estimation was used, the plotted response is on a logarithmic scale. If the subtracting model is accepted, then Stevens's graph implies that the true response measure is approximately equal to the logarithm of the magnitude estimate.

A series of papers by the Oregonians (e.g., Curtis, Attneave, & Harrington, 1968; Curtis & Fox, 1969; Fagot & Stewart, 1969) have studied sum and difference judgments using an iterative procedure to fit the equation,

$$R_{ij} = a(S_i^k \pm S_j^k)^m + b, \quad (6)$$

where R_{ij} is the response to the pair of physical stimuli, S_i and S_j , a and b are constants, and k and m are power function exponents. If the quantity in parentheses is interpreted as the sum or difference of the subjective values, then m can be interpreted as the exponent of a response output function. Values of m near 1 would indicate a valid, unbiased response measure.

The general outcome of these studies has been that m is near 1 when a category response is used. That agrees with the present emphasis on rating responses. On the other hand, m is greater than 1 for magnitude estimation. That is consistent with the hypothesis that magnitude estimation is increasingly distorted upwards as stimulus intensity increases.

Functional measurement procedure provides a superior method of analysis for sum and difference models than has been used by Curtis et al. With an interval response measure, both models imply parallelism, so the test can be made directly on the raw data. The assumption of power functions in Eq. 6 is unnecessary, and there are no parameters that need be estimated. A simple, powerful test of fit is available with analysis of variance. Stimulus scales are obtained directly from the marginal means. Furthermore, a nonlinear response measure will appear directly as nonparallelism, as in Weiss (1972).

These advantages of functional measurement procedure are important. The procedures used by Curtis and his associates seem adequate to indicate that ratings are superior to magnitude estimates. However, they are not sufficient to validate the rating scale. Indeed, the one paper that used an adequate test of goodness of fit found reliable discrepancies from the model (Fagot & Stewart, 1969). This is a matter of some concern since the present heaviness exponent of about unity differs substantially from the value of .65 obtained by Curtis et al (1968).

Finally, two discordant reports should be noted. Parducci, Thaler, and Anderson (1968) did not obtain parallelism for loudness averaging, but rather a convergence interaction. Birnbaum, Parducci, and Gifford (1971) obtained similar results for length averaging. Neither report, unfortunately, employed two precautions that have been standard in the present program of research, namely, stimulus end anchors and a moderately large number of steps in the rating scale. Both precautions are intended to minimize end effects and nonlinearities in scale usage. Despite these methodological questions, the discrepancy may be real. Unpublished

data by Birnbaum (personal communication, 1972) show a considerably smaller but still significant interaction when a graphic rating is used in the length-averaging task.

The averaging model, it should be noted, will produce an interaction when the stimuli within a factor have unequal weight. Indeed, the cited convergence interaction is what would be expected if the more intense stimuli also had greater weight. Such differential weighting has been obtained with verbal stimuli (Anderson, 1972), but whether the same holds for psychophysical stimuli is not clear. A direct test could be made if some way were found to manipulate the weight parameter without affecting scale value. In the loudness averaging task, for example, that might be done by varying the duration of the sounds.

The Psychophysical Law

In the functional measurement approach, the psychophysical law is derivative from the more basic psychological law. The traditional orientation, of Stevens and Fechner both, has placed primary emphasis on establishing the psychophysical law. Their work has been controversial because it failed to provide a validating criterion. And, from the present standpoint, their orientation has been misdirected. Primary emphasis should be on the psychological law; the psychophysical law is derivative from that.

Essentially, this point can be seen in many of the articles that have questioned magnitude estimation (see references in Anderson, 1970a, p. 166). Ratings vs magnitude estimations is not the central issue, but rather the need for a validating criterion. A psychological law provides such a criterion. It gives a validation base for the subjective scale of the stimulus.

Once s is known, determination of the psychophysical law, at least in its numerical form, is straightforward. That was illustrated in Fig. 3, which plotted s against S . That plot is the psychophysical law.

Some uncertainty still remains, of course, since various mathematical functions could do equally well with a finite set of imperfect data. Power functions may be quite useful for this purpose. The bias in magnitude estimation need not rule out power functions. Indeed, Marks (1968) has claimed that power functions will fit ordinary rating data. However, the usefulness of power functions probably reflects their flexibility in fitting a simple class of monotone functions (Mandel, 1964), not any

deeper psychophysical truth as has been argued by Stevens (1957) and Guilford (1932, 1954).

The Psychological Law

The psychological law has deeper significance than the psychophysical law. It is the psychological law that deals with the substantive problems of stimulus integration. At the same time, it provides a basis for scaling stimuli and response. Substantive theory and measurement theory are organically related and cofunctional in development.

The concept of a measurement scale carries with it a requirement of generality. The same scale should appear in various situations. This requirement is obvious (Anderson, 1962, p. 410), but three interwoven aspects deserve brief comment. These three aspects pertain to the stimulus scale, the response scale, and the psychological law.

Stimulus. The present data show the same stimulus scale to be operative in two different judgment tasks. An extremely interesting parallel, using averaging, difference, and ratio judgments with lifted weights, has been given by Birnbaum and Veit (1972). Such cross-task agreement naturally supports the validity of the stimulus scaling procedure. And since the stimulus scales are derivative from the response scales, the cross-task agreement also supports the validity of the response scale.

Response. On the response side, of course, the problem of generality is more extensive than on the stimulus side. The cumulation of evidence indicates that ratings can give interval scales across quite different stimulus situations.

Generality of the response scale has utmost importance. Simple algebraic models can be expected to hold only under restricted conditions. Stimulus interaction does occur and may be more the rule than the exception. With stimulus interaction, it becomes vital to be able to trust the response scale. A valid response scale allows a radical simplification of analysis and interpretation. This is one reason for stressing the need to develop numerical response measures.

As an incidental corollary to the above argument, it follows that nonmetric, ordinal methods have limited potential. To develop nonmetric methods that are practically effective for simple algebraic models may very well be possible. But to go beyond to study stimulus interaction in any general way with nonmetric methods seems out of the question.

Model. The final aspect of the problem of generality deals with the psychological law itself. To find the

same law in diverse judgment tasks, as with the averaging model (Anderson, 1972), is naturally desirable.

Such generality also bears on a problem of indeterminacy in the form of algebraic models (e.g., Anderson, 1970a, p. 161; Treisman, 1964; Zinnes, 1969). To illustrate, consider the simple averaging model. The parallelism of Fig. 1 is consistent with an arithmetic mean. An exponential transformation would change the arithmetic mean to a geometric mean; it would also change the parallel lines to a diverging fan of straight lines. The linear averaging model would then become a multiplying model, in which form it would be consistent with the transformed response. Since the exponential transformation is an allowable monotone transformation in functional measurement, there is a certain degree of arbitrariness in interpreting any single experiment. Each averaging model would be equally consistent with its own form of the data.

The interpretational problem is transfigured when diverse experiments agree in terms of scales and models. Two aspects may be briefly noted.

First, working within a coherent system may suffice even if there is an ultimate element of convention in the system. The cross-task coherence of results based on the arithmetic mean, for example, would be invariant under exponential transformation.

Second, and more important, certain judgment tasks impose plausible constraints. It may be reasonable to consider that Ss average lifted weights by geometric means, but it is not reasonable to suppose that they do the same with two-digit numbers, with spatial position, or with personality traits of positive and negative value. Similarly, an exponential transformation of a multiplying model (Anderson & Weiss, 1971) or a multilinear model (Shanteau & Anderson, 1972) is unattractive, because it does not produce a simple model.

The operative principles in this matter are the well-known ones of simplicity and unity. The same integration processes should be at work in diverse judgment tasks. Only with a widespread network of empirical evidence can a satisfactory resolution be reached. Problems of stimulus integration thus deserve more concentrated and wider study, for the above reasons and for their own interest as well.

REFERENCES

- ANDERSON, N. H. On the quantification of Miller's conflict theory. *Psychological Review*, 1962, 69, 400-414.
- ANDERSON, N. H. Application of a weighted average model to a psychophysical averaging task. *Psychonomic Science*, 1967, 8, 227-228.
- ANDERSON, N. H. Functional measurement and psychophysical judgment. *Psychological Review*, 1970a, 77, 153-170.
- ANDERSON, N. H. Averaging model applied to the size-weight illusion. *Perception & Psychophysics*, 1970b, 8, 1-4.
- ANDERSON, N. H. Test of adaptation-level theory as an explanation of a recency effect in psychophysical integration. *Journal of Experimental Psychology*, 1971, 87, 57-63.
- ANDERSON, N. H. Information integration theory: A brief survey. Technical Report No. 24, Center for Human Information Processing, University of California, San Diego, April 1972.
- ANDERSON, N. H., & JACOBSON, A. Further data on a weighted average model for judgment in a lifted weight task. *Perception & Psychophysics*, 1968, 4, 81-84.
- ANDERSON, N. H., & WEISS, D. J. Test of a multiplying model for estimated area of rectangles. *American Journal of Psychology*, 1971, 84, 543-548.
- ATTNEAVE, F. Perception and related areas. In S. Koch (Ed.), *Psychology: A study of a science*. Vol. 4. New York: McGraw-Hill, 1962.
- BAKER, K. E., & DUDEK, F. J. Weight scales from ratio judgments and comparisons of existent weight scales. *Journal of Experimental Psychology*, 1955, 50, 293-308.
- BECK, J., & SHAW, W. A. Ratio-estimations of loudness-intervals. *American Journal of Psychology*, 1967, 80, 59-65.
- BIRNBAUM, M. H., PARDUCCI, A., & GIFFORD, R. K. Contextual effects in information integration. *Journal of Experimental Psychology*, 1971, 88, 149-157.
- BIRNBAUM, M. H., & VEIT, C. T. Psychophysical measurement: Convergence criterion for rescaling. Paper presented at the meetings of the Western Psychological Association, Portland, April 1972.
- CURTIS, D. W., ATTNEAVE, F., & HARRINGTON, T. A test of a two-stage model of magnitude judgment. *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.
- EKMANN, G. Two generalized ratio scaling methods. *Journal of Psychology*, 1958, 45, 287-295.
- FAGOT, R. F., & STEWART, M. Tests of product and additive scaling axioms. *Perception & Psychophysics*, 1969, 5, 117-123.
- GARNER, W. R. A technique and a scale for loudness measurement. *Journal of the Acoustical Society of America*, 1954, 26, 73-88.
- GOUDE, G. *On fundamental measurement in psychology*. Stockholm: Almqvist & Wiksell, 1962.
- GUILFORD, J. P. A generalized psychophysical law. *Psychological Review*, 1932, 39, 73-85.
- GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.
- GUILFORD, J. P., & DINGMAN, H. F. A validation study of ratio-judgment methods. *American Journal of Psychology*, 1954, 67, 395-410.
- HARPER, R. S., & STEVENS, S. S. A psychological scale of weight and a formula for its derivation. *American Journal of Psychology*, 1948, 61, 343-351.
- MANDEL, J. *The statistical analysis of experimental data*. New York: Wiley, 1964.

- MARKS, L. E. Stimulus-range, number of categories, and form of the category scale. *American Journal of Psychology*, 1968, 81, 467-479.
- OTTANDER, C. *On sensory interaction in judgment of velocity and its measurement by psychophysical scaling methods*. Stockholm: Ivar Haeggström, 1967.
- 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.
- SHANTEAU, J. C., & ANDERSON, N. H. Integration theory applied to judgments of the value of information. *Journal of Experimental Psychology*, 1972, 92, 266-275.
- STEVENS, J. C., & RUBIN, L. L. Psychophysical scales of apparent heaviness and the size-weight illusion. *Perception & Psychophysics*, 1970, 8, 225-230.
- 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. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, 54, 377-411.
- TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.
- TREISMAN, M. Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology*, 1964, 16, 11-22.
- WARREN, R. M., & WARREN, R. P. Effect of the relative volume of standard and comparison-object on half-heaviness judgments. *American Journal of Psychology*, 1956, 69, 640-643.
- WEISS, D. J. Averaging: An empirical validity criterion for magnitude estimation. *Perception & Psychophysics*, 1972, 12, 385-388.
- WEISS, D. J., & ANDERSON, N. H. Subjective averaging of length with serial presentation. *Journal of Experimental Psychology*, 1969, 82, 52-63.
- WEISS, D. J., & ANDERSON, N. H. Use of rank order data in functional measurement. *Psychological Bulletin*, 1972, 78, 64-69.

NOTE

1. Weiss, D. J. FUNPOT, a FORTRAN program for finding a polynomial transformation to reduce any source of variance in a factorial design. Unpublished paper, June 1972.

(Received for publication July 5, 1972.)