

A test of a two-stage model of magnitude judgment¹

DWIGHT W. CURTIS, CALIFORNIA STATE COLLEGE, FULLERTON
 FRED ATTNEAVE AND THOMAS L. HARRINGTON, UNIVERSITY OF OREGON

It has been suggested that the power law $J = a^n$, describing the relationship between numerical magnitude judgments and physical magnitudes, confounds a sensory or input function with an output function having to do with O's use of numbers. Judged magnitudes of differences between stimuli offer some opportunity for separating these functions. We obtained magnitude judgments of differences between paired weights, as well as magnitude judgments of the weights making up the pairs. From the former we calculated simultaneously an input exponent and an output exponent, working upon Attneave's assumption that both transformations are describable as power functions. The inferred input and output functions, in combination, closely predict the judgments of individual weights by the same Os. Although pooled data (geometric means of judgments) conform fairly well to a linear output function, individual data do not; i.e., individual Os deviate quite significantly from linearity and from one another in their use of numbers. Individual values of the inferred sensory exponent, k , show significantly better uniformity over Os than do values of the phenotypical magnitude exponent previously found to describe interval judgments of weight.

Stevens (1957) has convincingly demonstrated that direct magnitude estimations are approximately related to values on a physical stimulus continuum by a power function of the form

$$J = a\phi^n. \quad (1)$$

In this equation, J is a numerical estimate of the magnitude of a stimulus with physical value ϕ , a is a constant which defines the scale unit, and the exponent n is a constant specific to the sensory modality. The function, with some minor elaborations, has been shown to hold for a wide variety of modalities. Unfortunately, it has been specific to this particular class of scaling procedures. Other experimental operations, particularly those assumed to produce interval scales, typically result in scales that are nonlinear with magnitude estimations.

These discrepancies among scales produced by different procedures have been variously interpreted. Stevens (1957) has hypothesized that scales based upon interval procedures are biased, because Os are unable in their judgments to avoid confounding the discriminability of stimuli with the perception of difference. Torgerson (1961) has proposed that a single quantitative relationship is perceived by the observer, which may be interpreted as a ratio or as a difference, according to the instructions

given by the experimenter. Garner (1953, 1954) and Attneave (1962) have each proposed that subjects are capable of evaluating the true relationships but experience difficulty in translating these relationships into numerical scale values. Attneave specifically proposed that the numerical estimation of magnitudes may reflect biases introduced in the mapping of sensory quantity on to a nonlinear continuum of numbers; i.e., the subjective number system may itself be a psychophysical function of its arithmetic counterpart. He suggested a two-stage model in which both the sensory input and the output of numerical judgments are described by separate power transformations. From several converging lines of evidence he inferred that the exponent for numbers would be of the order of 0.40.

Suppose that the psychological magnitude, ψ , is a power function of the stimulus, ϕ , such that

$$\psi = \phi^k. \quad (2)$$

Suppose further that the numerical judgment, J , is some function f of the psychological value. If f is also a power function, it follows that

$$J = a\psi^m. \quad (3)$$

The function describing the relationship between values on the stimulus continuum and responses would then be of the form

$$J = a'\phi^k m. \quad (4)$$

where k and m are the input and output parameters, respectively.

From the data of the usual magnitude estimation experiment it is not possible to separate the contributions of input and output transformations. We may consider another situation, however, in which subjects respond in a similar fashion to the differences between stimuli. In this instance the two transformations should be independent; the sensory exponent k should operate upon values on the stimulus continuum, and the "output" exponent m should operate upon differences. Hence the function describing the psychophysical relationship for differences should be of the form

$$J_{ij} = a(\phi_j^k - \phi_i^k)^m, \quad (j > i) \quad (5)$$

The foregoing relationships make possible the following analysis:

(1). We may estimate the values of the param-

eters k and m by performing a least squares analysis of the difference equation (5).

- (2). A second estimate of m may be obtained by a simultaneous consideration of the input parameter k and the exponent n of the power function (Equation (1)) fitted to magnitude estimation data. The parameter n in Equation (1), we have seen, should be equal to the product of k and m in Equation (5). (Bear in mind that km and n come from independent experimental operations.) Hence m may be independently estimated by

$$m = n/k \quad (6)$$

Using the value of n obtained in the least squares analysis of Equation (1), and k from the solution of (5), we may determine m . In the present experiment, this analysis was applied to judgments of lifted weights.

METHOD

Apparatus

The apparatus used for the presentation of stimuli was identical to that employed by Pradhan and Hoffman (1964). Subjects were seated in an arm-desk, the writing surface of which had been extended by an 18 x 24 in. plywood board. Two lengths of nylon fishline with tabs for lifting extended from holes in the top of the plywood surface. These lines passed through the plywood in flanged glass tubing to minimize friction. Beneath the plywood surface, attached to the nylon lines, were hooks on which the E suspended the weights to be lifted by the O. The arrangement permitted the quick interchange of weights and eliminated visual cues which might otherwise have been confounded with the sensation of weight.

A common set of eight weights was used in both experiments. These were spaced at approximately equal logarithmic intervals: 10, 16, 27, 44, 70, 116, 189, and 310 g. All 28 possible pairs of the eight weights were used in the difference estimation task.

Subjects

The 10 Os who participated in the main experiment were paid volunteers, undergraduate students at the University of Oregon and California State College at Fullerton. They were all naive with respect to psychophysical scaling.

Procedure

Each O estimated both differences and magnitudes. The difference estimation task was presented first.

The following instructions were given:

(1) *Magnitude estimations of differences.* "In this experiment, I would like you to estimate how much difference there is between some weights, when I present them to you in pairs. I will hang weights

on the end of the strings connected to these tabs, so that you can lift the weights by lifting the tabs. Begin by lifting the weight on the left, then the one on the right. You may lift the two weights as many times as is required to arrive at a judgment; but always begin with the one on your left and end with the one on your right. [Demonstrated] To estimate the difference between the two weights, I would like you to give me a number which represents how large the difference appears to you. For the first pair that I give you, the number that you assign will necessarily be arbitrary; use any number that seems appropriate, and comfortable to you. When you are presented with the second pair, and successive pairs, you should give me numbers which reflect how much difference there appears to be between the weights that you are currently evaluating, relative to the difference that you estimated immediately previously. That is to say, if you called the difference between two stimuli '27,' and the following stimuli seem to differ by about twice as much, you would give me a number about twice as large. If the second difference appears to be about ten times as great, give me a number like 270. If it appears to be one-fourth as great, give me a number like 6 or 7, etc. You may use any positive numbers which seem appropriate, including fractions or decimals. Try to make each of your judgments independent. That is, if a particular pair of weights seems very similar to a pair that you have previously lifted, don't try to remember the number that you assigned previously. Remember, for each trial, the present difference is judged relative to the immediately previous one. Are there any questions?"

(2) *Magnitude estimation of stimuli.* "In this task, some weights will be presented to you one at a time. I want you to judge how heavy each weight is by giving me a number which represents its 'heaviness.' Just as in the other task, the first number that you give me will be arbitrary. When subsequent stimuli are presented, however, you should assign numbers so that the numbers assigned reflect the magnitude of this stimulus relative to the magnitude of the immediately previous stimulus. That is, if one stimulus was assigned the value 27, and the next stimulus seems five times as heavy, you would say '130' or '140,' etc. It is important for your judgments to be independent, so that each represents your evaluation of the present stimulus relative to the immediately previous stimulus. Don't try to be consistent in the numbers that you apply to particular stimuli; if you called a particular stimulus a particular number on a previous trial, but a different number seems appropriate now, give me the latter number. Are there any questions?"²

To estimate differences, O lifted each of a pair of weights suspended by hooks at the ends of the nylon cords by pulling upward on the cloth tabs.

He was permitted to lift a pair of weights as many times as was necessary to arrive at a judgment, but was required always to begin with the right-most weight and end with the left one. This was to insure that the weights would be lifted an equal number of times. Each of the 28 possible pairs was presented five times. Both order and right-left position of presentation was randomized.

In estimating magnitude, a single weight was presented on each trial. The eight stimuli were presented in random order 10 times each. Ten practice trials were administered prior to each condition to acquaint O with the requirements of the task.

RESULTS

The empirical relationship between weights and numerical judgments was found to deviate from a simple power function at small stimulus values. Similar deviations have previously been found by Ekman (1961), McGill (1960), and Stevens (1961). Often an additive constant has been incorporated into the power function to correct for such deviations.

Functions incorporating the constant have been of two forms, representing alternative hypotheses concerning the significance of the constant. Ekman and Stevens have performed the correction upon the stimulus scale, justifying their procedure by postulating a relationship of the constant to the absolute threshold. McGill, however, has preferred to apply the correction to responses, assuming that the deviations reflect constant errors of judgment which represent, in effect, a displacement of the origin on the response scale. The latter of the two interpretations has been followed in the present analysis.

If McGill's assumption holds, the function for describing the relationship of magnitude judgments to physical values takes the form

$$J = a\phi^n + b. \quad (7)$$

where b constitutes the additive constant on the response scale. A generalization of this function to difference estimation would be

$$J_{ij} = c(\phi_j^k - \phi_i^k)^m + d. \quad (8)$$

The values of the parameters of each best fitting equation were estimated by a least squares curve-fitting procedure, programmed for the computer. This analysis was performed upon the judgment scales of each O individually, and upon a group scale obtained by pooling across Os. J values are geometric means of judgments at each stimulus value. In constructing the group scale, the judgments of each O were weighted by an appropriate factor to correct for individual differences in scale units.

Pooled data

Figure 1 shows the relationship between the geometric means of the pooled difference estimates and stimulus differences. The least squares description of the pooled judgments in terms of Equation (8) is

$$J_{ij} = 2.252(\phi_j^{.645} - \phi_i^{.645})^{1.141} - .671.$$

The fit of the function to the data is excellent, as shown in Fig 2; let us hold in abeyance the question of whether this is really any better than a function assuming linearity of output.

The relationship between the pooled magnitude estimation scale and stimulus values is described by

$$J = 1.722 \phi^{.746} - 4.852.$$

The assumption that the exponents k and m constitute input and output parameters, respectively,

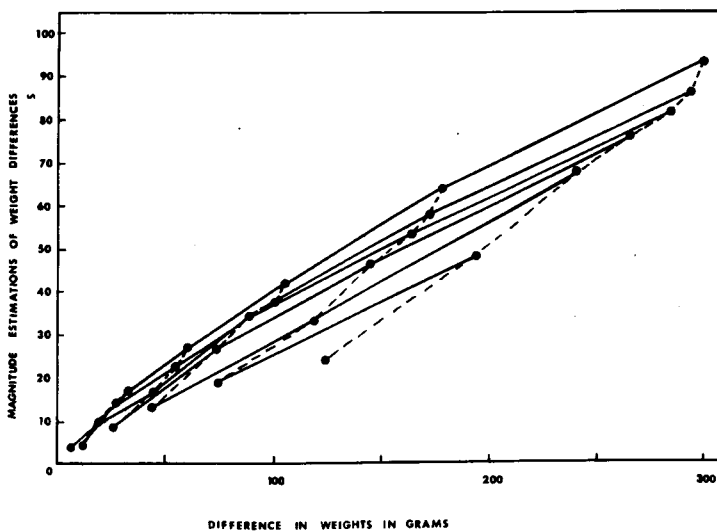
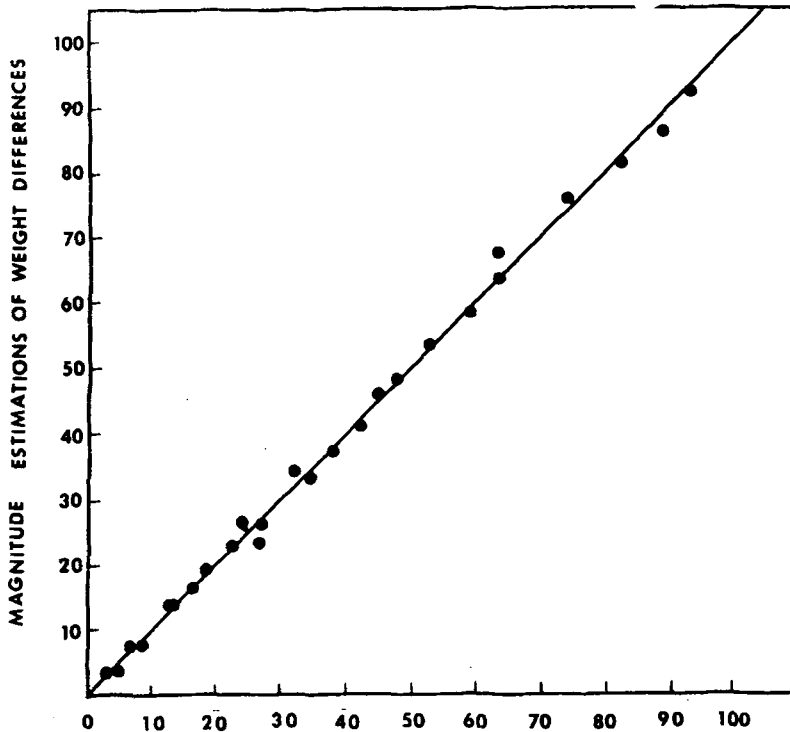


Fig. 1. Magnitude estimations of weight differences between weights. Solid lines connect responses to stimulus pairs for which the smaller of the two stimuli is common. The dotted lines connect responses to pairs for which the larger stimulus is common.



VALUES OF THE DIFFERENCE FUNCTION:

$$J_{ij} = 2.252 (\phi_j^{.645} - \phi_i^{.645})^{1.141} - .671$$

Fig. 2. Magnitude estimations of weight differences as a function of values on the difference model.

which characterize weight estimation as well as difference estimation, requires that the independent operations proposed in the introductory section yield reasonably similar estimates of m . The convergence of these estimates is in fact excellent. The relationship $m=n/k$ yields an estimate of $m=1.155$, which compares very favorably with the value of 1.141 directly estimated by a least squares analysis of Equation (8). This is to say that two mathema-

tically independent operations on the same sample of Os yield convergent values of the average output exponent; how well this value represents Os more generally is a question that remains to be considered. Alternatively, the product of k and m from Equation (8) ($.645 \times 1.141 = .736$) closely predicts the value of $n=.746$ obtained from judgments of individual weights.

In Fig. 3a we see the closeness with which difference judgments of the various pairs of weights are predicted from magnitude judgments of the weights making up the pairs when both are corrected for the nonlinearity of the output function. The corresponding predictions without correction are shown in Fig. 3b; the fit is still very good, but not quite as close as in 3a.

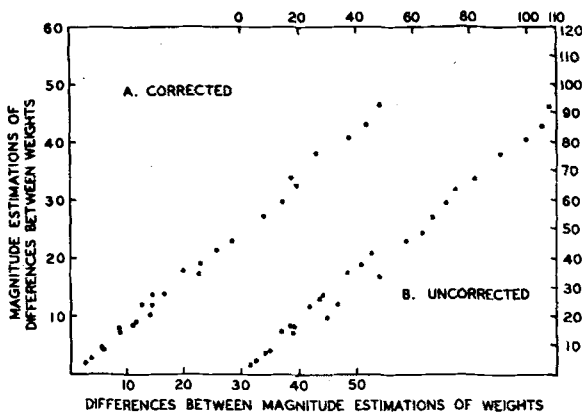


Fig. 3. Magnitude estimations of weight differences as a function of differences between magnitude estimations of weights. In 3A the means of each of the pooled judgments have been "corrected" by raising them to the power $1/m$ ($1/1.141 = 0.876$). The effect of this treatment may be evaluated by comparing the function with the one in 3B, in which mean judgments are not corrected.

Individual differences

The exponents of individual Os for Equations (7) and (8) are presented in Table 1, with associated standard errors. These standard errors merely describe the uncertainty with which the parameters are determinable from O's own judgments (or from the geometric means of all these particular Os' judgments, in the case of pooled data), and imply nothing about generalization to a parent population. The bottom line of Table 1 shows parameter values averaged over Os: These should not and do not differ to any important degree from corresponding pooled-scale values, since they are based on exactly the

Table 1. Parameter Values of Magnitude and Difference Functions: Individual, Pooled, and Mean Scales

	Parameters			
	$J = a\phi^n + b$ n	$J_{ij} = a(\phi_j^k - \phi_i^k)^m + b$ k	m	$m = n/k$
Subject				
K. J.	.600 ± .135	.652 ± .048	.917 ± .077	.920
R. F.	.583 ± .065	.676 ± .141	1.019 ± .147	.862
J. W.	.562 ± .076	.673 ± .035	.873 ± .074	.836
W. H.	.702 ± .130	.672 ± .058	.998 ± .088	1.044
C. L.	.603 ± .130	.546 ± .062	1.055 ± .120	1.106
L. G.	.691 ± .121	.584 ± .053	1.007 ± .098	1.183
L. K.	.392 ± .080	.494 ± .065	.714 ± .115	.793
D. E.	.969 ± .069	.759 ± .073	1.432 ± .166	1.277
C. F.	.904 ± .114	.477 ± .052	1.919 ± .187	1.895
B. R.	1.043 ± .117	.785 ± .087	1.402 ± .190	1.328
Pooled Scale	.746 ±	.645 ± .020	1.141 ± .033	1.155
Parameter Means with between-O Standard Errors	.705 ± .065	.632 ± .033	1.134 ± .112	1.124 ± .104

same data averaged in one case before curve-fitting, in the other after curve-fitting. The standard errors on the bottom line are based on *between-O* variability; it is these that must be considered if we are interested in generalizations concerning population parameters.

Note, in the first place, that the variability of Os (directly proportional to between-O standard error) is only about half as great on k as on n. This difference is significant at the .05 level ($t=2.56$, $df=8$). (The test for a difference between correlated variances is described by McNemar, 1962.) This is exactly the sort of result that is to be expected if most of the disagreement between Os on the exponent of the direct magnitude judgment function, n, is attributable to differences between their output functions.

Turning to the output exponent, m, we see that the value for the pooled judgments, 1.141, deviates from unity by more than four times its standard error, .033, indicating that the average output function is reliably nonlinear for this particular sample of Os. On the other hand, the deviation from unity is small in reference to the between-O standard error, .112; thus our data in no way disconfirm the hypothesis that, over some large parent population of Os, the average value of m might be one. We state the matter in this somewhat cumbersome way because it is necessary to distinguish carefully between two possibilities (actually ends of a continuum of possibilities). One is that there is a "true" parameter value (e.g., $m=1$) common to everybody, and that combining data from different people averages out random error and thereby improves our estimate of this true value. The other is that people are genuinely and reliably different on the parameter, in which case averaging their data does not eliminate error, but merely provides a summary statistic of limited psychological interpretability.

The strongest evidence that our Os differ reliably on m is presented in Fig. 4. Here the two independent estimates of the output exponent are plotted one against the other for each O ($r=.96$, $p<.001$). The vertical lines extend 1 SE (within individual O) above and below each point; since only one point deviates from the equality line by more than one SE, it is obvious that the correlation is about as good as the reliability of the values permits.

A corresponding comparison of k with $k=n/m$ yields $r=.80$. This is by no means independent of the m vs m relationship; the sources of residual error are approximately the same in both cases; the essential reason for the differences in rs is that marginal variability is greater on m than on k (for m, $s=.353$; for k, $s=.104$).

It is abundantly evident from these data that people are abundantly different in their psychophysical judgments. That such nonrandom differences are predominantly on the output rather than the input side is strongly suggested by the fact that variance over Os is four times as great on n as on k, and by the differences between correlation coefficients and marginal dispersions just noted.

There is further some suggestion that within-O error on m may be contributing at least a little to apparent individual differences on k. In Fig. 5 the standard error of m is plotted against k for each O: From this it seems that the calculated value of k is most likely to be deviant from the group average in those cases where the true value of m is least determinate, and we are accordingly most uncertain what the O's output function is and how to correct it. A similar comparison of k values with their own, rather smaller standard errors does not suggest any such relationship.

DISCUSSION

For the moment let us forget about individual

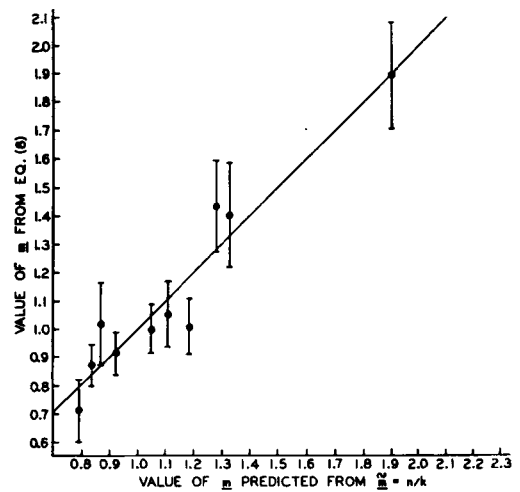


Fig. 4. Relationship between values of the output exponent m and values of the same exponent predicted by the relation $m = n/k$ ($r = .96$, $p < .001$).

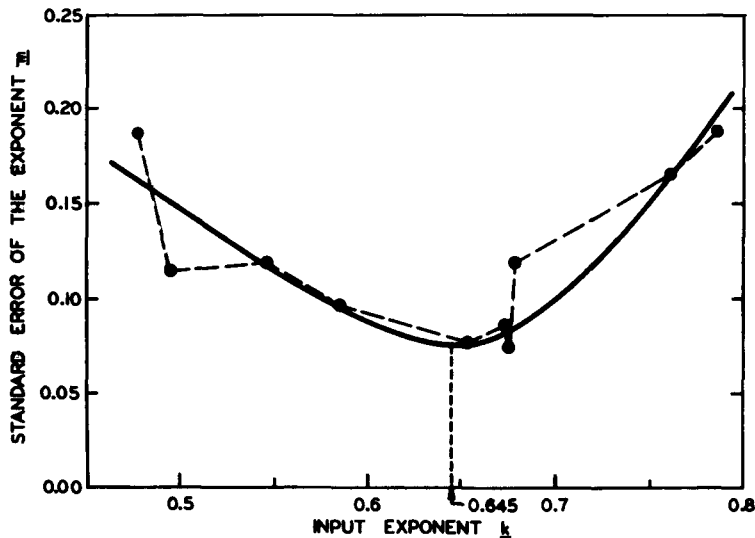


Fig. 5. Standard errors of m for the difference functions of individual Os as a function of k .

differences and pretend that people are all alike, aside from random error. Considered on this basis, our group data are reasonably consistent with Stevens' assumption that people use numbers in a linear manner. On the other hand, our results are strongly in disagreement with Attneave's earlier suggestion that the exponent for the number continuum has a value of about .4 (i.e., that $1/m = .4$, or $m = 2.5$). They seem no less in disagreement with quantitative predictions of other multistage models—see Torgerson (1961), Ekman (1964), and Treisman (1964)—and with empirical results from at least two other modalities.

Torgerson (1961) obtained magnitude judgments of differences between stimuli varying in lightness. If we interpret his brief summary correctly, these were approximately predictable from magnitude judgments of the individual stimuli by the relationship

$$\log J_{ij} = c(\log J_j - \log J_i).$$

This is the boundary condition of

$$J_i^{1/m} = c'(J_j^{1/m} - J_i^{1/m})$$

as $1/m$ approaches zero.³ It is consistent with the hypothesis of an output function even more deviant from linearity than that suggested by Attneave. Beck and Shaw (1966) have obtained similar results in a study of magnitude judgments of loudness differences.

Our group data seem peculiar in another way: They yield a phenotypical exponent of .75 for the power relationship between magnitude judgments and individual weights, in contrast to the exponents considerably greater than 1 that have been found in most earlier experiments, summarized by Stevens and Galanter (1957).

Now, the most plausible hypothesis we can suggest to account for these disagreements is that our Os deviated far less from linearity in their use of numbers than Os in other experiments have typically

deviated. (Note well that the Es who have reported weight exponents of 1.5 or thereabout had no way whatever of checking on response linearity, whereas we do have such a check.) It is hard to explain why fortuitous differences in experimental situations would affect response functions so drastically, but it is much harder to imagine that they would affect sensations of weight per se. Torgerson's (1960, 1961) basic argument is that people map their sensations upon the number continuum in different ways under different conditions; that the same subjective relation may be given either a "difference interpretation" or a "ratio interpretation," depending (usually) on whether interval or magnitude judgments are requested. Beck and Shaw (1961) independently developed the same idea, and showed that subjects judging pitch magnitudes could be influenced toward representing octaves either by equal differences or by equal ratios, depending on the position of the standard. We believe that this line of reasoning is essentially correct, with the provision that sensations seem to be mapped on to numbers in many ways, not just two. Attneave's (1962) implicit assumption of a single, invariant output transformation should certainly be abandoned in the light of current evidence.

Despite these disagreements with earlier results, we find one very striking convergence. For the average input function, we obtain an exponent of .645 from pooled data, or of .632 (SE = .033) by taking the mean of the parameter over Os. These estimates of an average k are almost identical with the exponent of .62 which Attneave (1962) estimated earlier from a graph of Cowdrick's interval-judgment data. Thus the hope that a two-stage model may reconcile differences between magnitude and interval scales remains very much alive.

It was certainly not our initial anticipation that individual difference data would provide the major

support for a two-stage model in this study. Such is the case, however, and the support turns out to be very powerful indeed. Anyone who wishes to maintain that numerical responses directly reflect sensory magnitudes is faced with a real challenge in Fig. 4 and similar evidence demonstrating the reliability of differences between individual output functions.

We have presented evidence that people differ predominantly on m rather than on k ; however, there is no reason why we should not expect small differences in input functions. It is well to remember that no single sensory transducer is responsible for weight sensations in a lifting experiment: Receptors in the wrist, the elbow, and several different muscle groups are involved; different Os might employ different lifting movements, their own hands and arms may differ in weight, and so on. In view of all these considerations, the dispersion of Os on k seem remarkably small.

In summary, we suggest that the following facts argue strongly for a two-stage interpretation of the present data: (1) Os are significantly more alike on the supposedly genotypical sensory exponent, k , than on the phenotypical exponent n . (2) Estimates of individual output exponents by two logically independent operations are highly convergent; $r = .96$ over Os. (3) Similar independent estimates of m from pooled data are even closer (1.141 and 1.155). (4) The average value of the sensory exponent, k , determined in accordance with the model, is almost identical with the exponent of a power function describing earlier interval-judgment data.

References

- Attneave, F. Perception and related areas. In S. Koch (Ed.), *Psychology: A study of a science*, Vol. 4, New York, 1962.
- Beck, J., & Shaw, W. A. The scaling of pitch. *Amer. J. Psychol.*, 1961, 74, 242-251.
- Beck, J., & Shaw, W. A. Ratio estimations of loudness intervals. *Amer. J. Psychol.*, 1966, in press.
- Ekman, G. Some aspects of psychophysical research. In W. A. Rosenblith (Ed.), *Sensory communication: Contributions to the Symposium on Principles of Sensory Communication*. M.I.T., Wiley, 1961.
- Ekman, G. Is the power law a special case of Fechner's Law? *Percept. mot. Skills*, 1964, 19, 730.
- Fagot, R. On the psychophysical law and estimation procedures in psychophysical scaling. *Psychometrika*, 1963, 28, 145-160.
- Garner, W. R. Context effects and the validity of loudness scales. *J. exp. Psychol.*, 1954, 48, 218-224.
- Garner, W. R. A technique and a scale for loudness measurement. *J. Acoust. Soc. Amer.*, 1954, 26, 73-88.
- McGill, W. The slope of the loudness function. In H. Gulliksem & S. Messick (Eds.), *Psychological scaling: Theory and applications*. New York: Wiley, 1960.
- McNemar, Q. *Psychological statistics*. New York: Wiley, 1962. P. 246.
- Pradhan, P. L., & Hoffman, P. J. Effects of spacing and range of stimuli in magnitude estimation judgments. *J. exp. Psychol.*, 1963, 66, 533-541.
- Stevens, S. S. On the psychophysical law. *Psychol. Rev.*, 1957, 64, 153-181.
- Stevens, S. S. The psychophysics of sensory function. In W. S. Rosenblith (Ed.), *Sensory communication: Contributions to the Symposium on Principles of Sensory Communication*. M.I.T., Wiley, 1961.
- Stevens, S. S., & Galanter, E. H. Ratio scales and category scales for a dozen perceptual continua. *J. exp. Psychol.*, 1957, 54, 377-411.
- Torgerson, W. S. Distances and ratios in psychophysical scaling. *Acta Psychol.*, 1961, 19, 201-205.
- Treisman, M. Sensory scaling and the psychophysical law. *Quart. J. Psychol.*, 1964, 16, 11-22.

Notes

1. This research was supported in part by NIH grant MH 14116 and in part by the Air Force Office of Scientific Research, Grant No. 973-66.
2. These instructions were evolved through several stages of trial and error. Preliminary Os, who were instructed in a more casual manner, tended to give judgments considerably less amenable to economical description in terms of either the present model or any other. Ceiling effects were particularly marked; i.e., O often seemed reluctant to give judgments greater than some arbitrary value as differences in magnitudes increased. We were faced with somewhat of a dilemma over whether or not to include the last 4 preliminary Os in the data analysis; these Os received almost the final instructions, the chief difference being that the tactic of making each judgment comparable to that of the immediately preceding trial was not explicitly emphasized. Our final decision was to exclude them, and use only the 10 Os who were instructed in the manner described above. If they had been included, no important conclusions of the study would be changed, but the data would be less orderly. These 4 Os showed relatively poor convergence between alternative determinations of m , and yielded highly variable values of k : .296, .428, .695, and 1.002 (cf. Table 1). The two extreme values are essentially uninterpretable, being based on data that are inconsistent with anything like a power model: in one case a strong ceiling effect is evident, in the other a peculiar plateau in the mid-region of response. In view of the subsequent stabilization of results, it seems most likely that these Os were deviant because of inadequate instructions; however we have no way of being really sure that they would have performed differently under the final instructions.
3. See Fagot (1963). Addition of a small constant to the J_{ij} 's and to the J_i 's respectively, before taking logarithms, would probably make for a better fit to Torgerson's data, since he says that the simpler relationship held "except at the bottom of the scale." Likewise, to make the power form above entirely consistent with the model we are currently testing would involve adding similar constants ($-b$ and $-d$ of equations 7 and 8) to the J 's before raising them to the $1/m$ power. Fagot shows that the convergence of the power function upon the log function as the exponent decreases still holds when such additive constants are included in both. It seems evident, therefore, that the power model would have to fit like those of Torgerson (1961) or Beck and Shaw (1966) at least as well as a logarithmic formulation.

(Accepted for publication September 12, 1967.)