Perception & Psychophysics
1977, Vol. 21 (3), 201-215

Note on functional measurement and data analysis

NORMAN H. ANDERSON
University of California, San Diego, La Jolla, California 92093

The guiding idea of functional measurement is that measurement theory and substantive theory form an organic unity. Psychological scales are inherent in the statement of quantitative psychological laws, and these laws themselves form the base and frame for psychological measurement. Valid scales thus depend on empirically valid laws. But establishing empirical validity of any law requires appropriate data analysis. Several statistical problems are discussed with respect to simple algebraic laws. To illustrate the necessity for proper tests of goodness of fit for algebraic models, five sets of experimental data are reanalyzed. In each case, the factorial plot and the analysis of variance showed that the data were nonadditive. Nevertheless, an additive model was fit to the data. The correlations between the data and the predictions from the additive model were extremely high, ranging from .964 to .9997. The corresponding observed-predicted scatterplots also gave little sign of the deviations from additivity. These correlation-scatterplot analyses conceal and obscure what the factorial plot and the analysis of variance reveal and make clear. Other topics discussed are accepting and rejecting the null hypothesis, the use of nonmetric smoothing for parameter estimation, and problems of stimulus-response-model generality. An extension of functional measurement is suggested for a practicable error theory for nonmetric analysis.

In the functional measurement approach, the essential element is the stimulus integration function that governs the synthesis of separate stimulus components into a unitary perception or judgment. This problem of stimulus integration ordinarily has primary substantive interest. However, the integration function also serves as the base and frame for constructing valid scales of subjective magnitude. The term "functional" reflects this basic conceptual and technical role of the integration function.

The guiding idea of functional measurement is that measurement theory and substantive theory form an organic unity (Anderson, 1962b, 1970). Psychological scales are derivative from the integration function. Applicability of functional measurement thus rests squarely on the empirical validity of the stimulus integration function. It is not enough to assume that some function is correct; that must be established empirically.

In principle, validation of an integration function is straightforward. It requires a test of goodness of fit, that is, an assessment of the deviations between the model and the data. In practice, unfortunately, this simple precept presents numerous difficulties. Many reports employ tests that are incorrect, or experimental designs that do not allow a valid test. And, at best, statistical analysis, although necessary, is not sufficient.

The purpose of this note is to comment on certain aspects of these questions. A preliminary overview will be given to clear up some apparent misunderstandings about the nature of functional measurement. Some problems of statistical analysis will then be discussed. In addition, a functional measurement approach will be given to an error theory for nonmetric analysis.

Preparation of this paper was supported by National Science Foundation Grant BMS 74-19124 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 Edward Carterette, Dwight Curtis, and David Weiss for their comments on an earlier draft. Thanks also to Cliff Butzin and Jim Zalinski for their help with data analysis.

THE LOGIC OF FUNCTIONAL MEASUREMENT

Functional Measurement Diagram

The nature of functional measurement can be exhibited most clearly in the functional measurement diagram (Anderson, 1970). A reduced version is given in Figure 1.

In the diagram of Figure 1, physical stimuli, S, impinge on the organism to be processed by the valuation function, V, into psychological stimuli, s. These psychological stimuli are combined by the integration function, I, into a covert response, r. This covert response is transformed by the response function, M, into the overt response, R.

In the case of simple sensory stimuli, the valuation function, V, is commonly called the psychophysical law. However, V is more general, applying also to verbal or symbolic stimuli that lack a physical metric.

The integration function, I, can also be called the psychological law, for it relates the internal stimuli

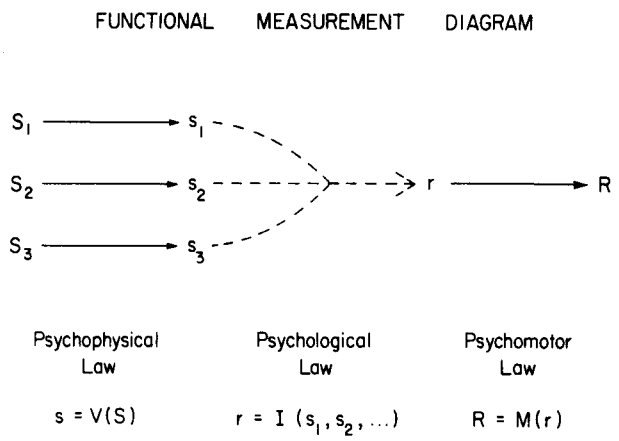

Figure 1. Functional measurement diagram.

to the internal response. This integration function constitutes the basis for functional measurement.

The response function, M, can also be called the output function or psychomotor law. It relates the covert response to the overt measurement scale imposed by the investigator. If M is linear, so that R is a linear function of r, then R is said to be a linear or interval scale (Anderson, 1975, p. 473).

This diagram shows a cascade of three conceptually distinct functions. If theoretical clarity is to be obtained, then the physical law that relates the observable stimuli and response must be broken down into these three component functions. A strict behavioristic view is not adequate (Anderson, 1974a, p. 284).

Two functions in Figure 1 are associated with problems of measurement. The V function refers to measurement of the stimulus, whereas the M function refers to measurement of the response. The solution to both measurement problems rests on the integration function, I.

The Functional Logic

A general theory of measurement must be able to work with ordinal or rank-order scales. A solution to this problem was given in the two papers that introduced the essential ideas of functional measurement (Anderson, 1962a, b). "The logic of the present scaling technique consists in using the postulated behavior laws to induce a scaling on the dependent variable" (Anderson, 1962b, p. 410). In terms of the present Figure 1, "behavior law" corresponds to integration function or psychological law, while "dependent variable" refers to the overt response scale, R.

If the integration function is empirically valid, then it can be established at the same time that the response scale is transformed into a linear measure. The fundamental element in this approach is the integration function; its mathematical form constitutes the essential basis for measurement, both of stimuli and of response. Related views have been expressed by Krantz (1972) and by Marks (1974, p. 277).

An essential property of the integration function is that it is a function of two or more stimulus variables. Consequently, monotone rescaling of the observable response measure still leaves degrees of freedom for testing the integration function, I (Anderson, 1962b, 1970, p. 167, 1975, p. 478, 1977).

This many-variable approach represents a distinct break from the one-variable approach that is typical of much work in psychophysics. The one-variable approach is inherently inadequate as a basis for measurement theory because it provides no protection from ordinal bias in the overt response scale. The many-variable approach can give adequate constraints on monotone response transformation, and so provides an adequate basis for theoretical development.

An Illustrative Model

For present illustration, suppose that the subject is asked to judge average or total intensity of two stimuli (Anderson, 1972, 1974c). Stimulus pairs are constructed according to an ordinary row-by-column factorial design. The natural model for this task is

$$r_{ij} = C_0 + s_{Ri} + s_{Cj}, \qquad (1)$$

where s_{Ri} and s_{Cj} are the subjective values of the row and column stimuli, C_0 is a zero-point constant, and r_{ij} is the covert response. (A complete statement of the model would include weight parameters for each stimulus dimension, as well as a term to represent response variability. Response variability will be explicitly considered below, but Equation 1 will suffice for present purposes.)

Case of Linear Response

When the overt response is a linear scale, then the analysis is very simple. This appears in the parallelism theorem which may be stated as follows:

If (1) the linear model is correct, and if
(2) the response measure is a linear scale, and if
(3) the stimulus variables have independent effects,
then (1) the data from a factorial design will plot as a set of parallel curves, and moreover,
(2) the marginal means of the data table will be the stimulus values on validated interval scales.

The proof of this theorem is straightforward and is not given here.

This theorem means that observed parallelism constitutes joint support for all three assumptions. This

follows because, if any one assumption fails, then parallelism will not in general obtain. There is, of course, a logical possibility that two assumptions might fail in a compensating way to produce parallelism.

This parallelism theorem has been extremely valuable in the development of information integration theory (Anderson, 1974a, b, d). The usefulness of the parallelism theorem results from two fortunate findings. The first, which has been widely doubted, is that ordinary rating methods can yield linear response scales with only modest experimental precautions (Anderson, 1974a, p. 231, 1974d, p. 245, Note 1). The second, which has been widely conjectured, is that stimulus integration obeys simple algebraic models in a variety of situations. Both findings have been placed on solid ground by the experimental work on information integration theory.

Weiss (1972) provides a relevant illustration of this point. His subjects judged average grayness of two Munsell chips using two response measures, ratings and magnitude estimation. Equation 1 is assumed to hold for the covert response. If either overt response is a linear scale, then the data will exhibit parallelism. Both measures have equal opportunity. Not both can succeed because, as is well known, they are nonlinearly related. As it happened, the rating method succeeded and magnitude estimation failed. Stevens' (1971) objection that the task itself imposes a bias is ad hoc and disagrees with much collateral evidence.

Case of Ordinal Response

If the overt response is only an ordinal scale, then parallelism will not obtain even though the model is true. This was the situation that was faced in the original article (Anderson, 1962b). The same problem arises in psychophysical bisection (Figures 3 and 4 below) where the overt response is in the physical metric.

In principle, however, the case of ordinal response is also straightforward. Since the response scale is ordinal, some monotone transformation will make it linear. This monotone transformation can be determined because it is the one that makes the data parallel, at least when the linear model is true.

In practice, of course, there are two difficult problems in the use of monotone transformations. The first is that of computing it. The original suggestion for the use of power series expansion (Anderson, 1962b) was developed statistically by Bogartz and Wackwitz (1971) and implemented in Weiss' (1973, 1975) FUNPOT program. This method provides a valid error theory. However, it seems to be less computationally desirable than nonmetric additivity programs such as MONANOVA (Kruskal, 1965) and ADDALS (de Leeuw, Young, & Takane, 1976).

The second major problem with using monotone transformation relates to two aspects of statistical procedure. One is the lack of an error theory, a necessity for model analysis. The other relates to inferential power. Ordinal, qualitative tests can be useful for eliminating hypotheses, as illustrated by the crossover interactions of averaging theory (Anderson, 1965, p. 397, 1974d, p. 249). However, monotone transformation clearly gives tremendous flexibility in fitting the data to the model. Good fits may be easy to obtain, even when the model is invalid (Anderson, 1962b, p. 410). Consequently, nonmetric analysis provides much less power than metric analysis for the fundamental problem of inverse inference, that is, from the overt responses to the underlying model. Both of these statistical problems are discussed below in the section on error theory for nonmetric analysis.

COMPARISON OF THREE METHODS FOR TESTING GOODNESS OF FIT

Functional measurement depends entirely on the validity of the integration function or model. Validity has to be established empirically, and that requires an assessment of how well the model fits the data. Unfortunately, some current methods for testing goodness of fit suffer from fundamental shortcomings. For example, the correlation between predicted and observed is not only logically inappropriate, but can be extremely high even though the model is seriously incorrect. To apply functional measurement, therefore, it is necessary to give adequate attention to the problem of testing the validity of the integration model.

Factorial Plots Compared with Correlations and Scatterplots

Five experimental examples will be used to compare three methods of testing goodness of fit. In each example, the original analysis was done using factorial plots and analysis of variance with satisfactory results. All five examples were reanalyzed using correlations and scatterplots between predicted and observed. As will be seen, correlations and scatterplots are generally inadequate if not downright misleading.

Grayness averaging. Subjects judged average grayness of two Munsell chips using magnitude estimation. The theoretical hypothesis was that the averaging model was correct so that the above parallelism theorem would apply. However, it was expected that the data would not actually plot as parallel lines, owing to bias in magnitude estimation.

The factorial plot of the data is in the upper panel of Figure 2. Each curve represents one row of the 5 by 5 design, and the corresponding Munsell value

of the left chip is listed as curve parameter. Similarly, the horizontal axis lists the Munsell value of the right chip.

The five curves of Figure 2 are markedly nonparallel. This nonparallelism can best be appreciated by measuring the vertical spread from right to left; it varies by more than 2 to 1. This nonparallelism was statistically significant in the analysis of variance test. Clearly, these data are not additive.

Nevertheless, an additive model was fit to the data in the manner described just below. The lower panel of Figure 2 plots the predicted values as a function of the observed. The fit looks respectable, and the correlation is quite high, .983. Closer study shows systematic deviation from the diagonal line, but the correlation-scatterplot analyses give little inkling of the extreme nonparallelism in the upper panel.

The additive model was fit to the data of Figure 2

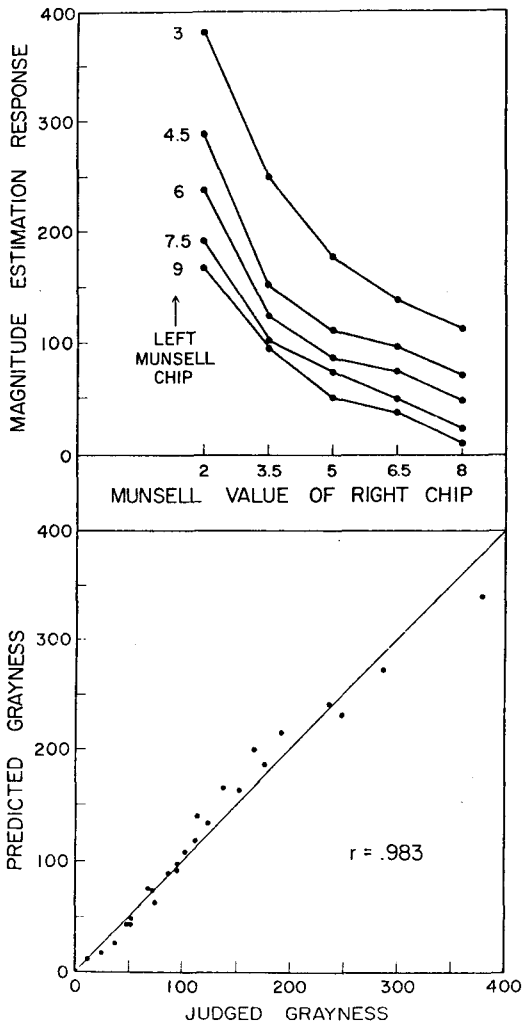

Figure 2. Upper panel gives factorial plot of mean judgment of average grayness as a function of Munsell value of left and right chips. Lower panel plots predictions from the additive model as a function of the observed means in the upper panel. (After Weiss, 1972.)

using a least squares criterion to estimate the stimulus values from the given data. Had the stimulus values been estimated separately from independent data, failure of the test of fit would have been ambiguous since it could be attributed to incorrect stimulus values. The least squares approach, which is equivalent to the analysis of variance model with no interaction term, gives the additive model its optimum possibility for fitting the data. Failure cannot then be attributed to invalid stimulus values per se. It may be added that this method also provides more power to detect discrepancies from the model because it avoids unreliability that would be introduced by the use of independent estimates of the stimulus values.

Grayness bisection. This experiment was designed to extend Weiss' (1975) functional measurement analysis of bisection, using a larger design for more power and for a better determination of the psychophysical law. Subjects selected a Munsell chip to lie halfway in grayness between two given chips. The two given chips were varied in factorial design, and the response measure was the Munsell value of the chip selected as the bisector.

The theoretical hypothesis is that the bisection task obeys a weighted average model (Anderson, 1970, Equation 9). If this model is true, and if the Munsell scale is a true linear or equal-interval scale of grayness, then the measured response data should have the additive form.

Accordingly, an additive model was fit to these data in the same way as above. The plot of predicted vs. observed is in the lower panel of Figure 3. The points cluster very closely around the diagonal line of perfect fit. The correlation between predicted and observed is exceptionally high, .997. An investigator who relied on such correlation-scatterplot statistics might readily conclude that the additive model held for bisection, and, simultaneously, that the Munsell scale was a true linear scale of grayness.

This conclusion is not justified, however, for the factorial plot in the upper panel of Figure 3 is nonparallel. The curves show a gradual rightward divergence that can best be appreciated by measuring the vertical spread between top and bottom curves. This vertical spread increases by 27% from left to right. This nonparallelism is reliable as shown by the interaction test of analysis of variance, $F(20,300) = 2.82$, $p < .0001$. These data, therefore, are nonadditive. More detailed analysis (Anderson, 1976b; see also below) shows that the cause of this failure of goodness of fit lies in the Munsell scale of grayness, not in the bisection model itself.

Length bisection. Six subjects were instructed to choose a variable rod so that its apparent length was halfway between two given lengths. A monotone transformation was applied to maximize additivity. If length bisection obeys the averaging model, then these transformed data will be additive.

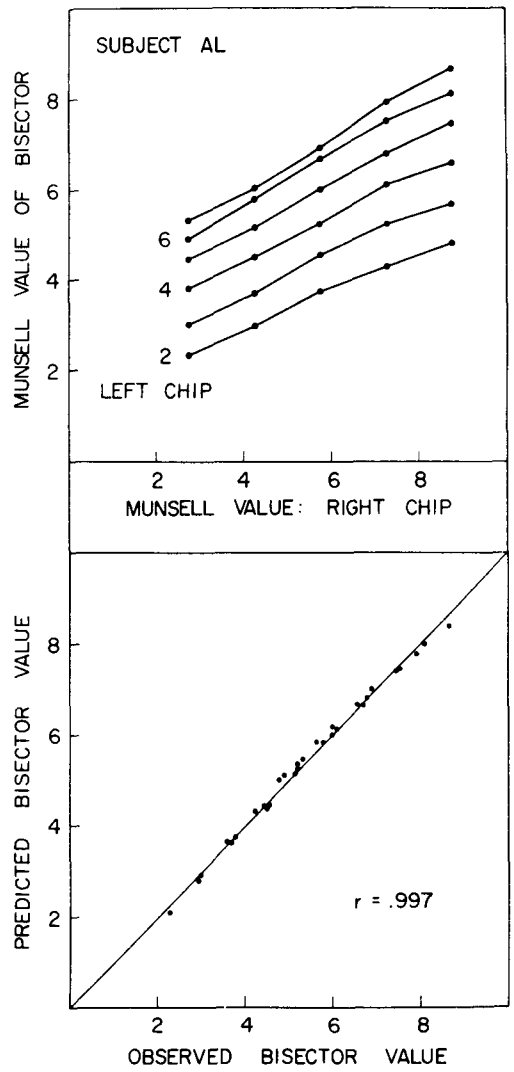

Figure 3. Upper panel gives factorial plot of mean Munsell value of bisector chip as a function of the Munsell values of the left and right bisectee chips. Lower panel plots predictions from the additive model as a function of the observed values in the upper panel. (Data from Subject A. L., after Anderson, 1976b).

The right panel of Figure 4 plots the transformed data as a function of the predictions from the additive model. The correlation is surpassingly high, .9997. Deviations from the line of perfect fit are miniscule. These correlation-scatterplot analyses seem to say that the bisection model holds for the dimension of length.

However, the factorial plot in the left panel of Figure 4 shows small, but systematic, convergence to the right, most readily appreciated by measuring vertical spread. This nonparallelism in the transformed data was statistically reliable ($p = .001$) by application of the error theory for nonmetric analysis given below. In addition, a metric analysis of the raw data showed that no monotone transformation could make the data additive. In contrast to grayness, therefore, the bisection model does not appear to hold for length.

Loudness averaging. Subjects judged average loudness of a sequence of four noise bursts combined in a four-way factorial design. All six interactions involving the noise at Serial Position 4 were significant, so the data do not exhibit parallelism. An illustrative two-way interaction is shown in the upper panel of Figure 5, which gives the judgment as a function of the noise intensity at the last two serial positions. The curves show marked convergence to the right, with a decrease in vertical spread from 1.30 to .86 scale steps. This interaction was significant, $F(9,63) = 5.82$, $p < .00001$.

Nevertheless, an additive model was fit to the data, averaged over Serial Position 1 in order to reduce the number of data points from 256 to 64. The predicted-observed plot is shown in the lower panel of Figure 5. The correlation is .964, and the mean absolute discrepancy is less than .16 scale steps. Although hindsight might suggest some discrepancy in the scatterplot, it would be hard to attribute significance or meaning to it.

Here, again, the factorial analysis provides a superior portrait of the data. It is especially useful for the multiple-stimulus experiments.

Size-weight illusion. A pound of lead feels heavier than a pound of feathers because the visual cues have a strong effect. The theoretical hypothesis in this experiment was that the visual and the kinesthetic cues are integrated by a linear, or additive, rule.

Subjects lifted a cylinder and rated its heaviness on a 1-20 scale. The factorial plot of the data is shown in the upper panel of Figure 6. The cylinders varied in height as listed on the horizontal axis, and in gram weight as listed by each curve. The upward slope of the curves means that the same gram weight feels heavier in a smaller cylinder.

The factorial plot looks nicely parallel, except for one point, at the right end of the lower curve. This one-point discrepancy was enough to cause a significant deviation from parallelism. Its interpretation is discussed below.

The lower panel of Figure 6 plots the predicted-observed scatterplot for the linear model. The points cluster closely around the diagonal line of errorless fit. The correlation is exceptionally high, .996. This scatterplot gives little sign of the one-point discrepancy. Nor does it give the sense of data reliability that reflects from the factorial plot.

An incidental but noteworthy aspect of these data is their illustration of the great power of the analysis of variance test. The one-point discrepancy, small and local though it is, produced a significant result in the overall interaction test.

Comment on Correlations and Similar Statistics

How can invalid models produce such high correlations? The answer is simple. Any model that is at all plausible will predict a low response to low stimuli, a high response to high stimuli. That is enough to

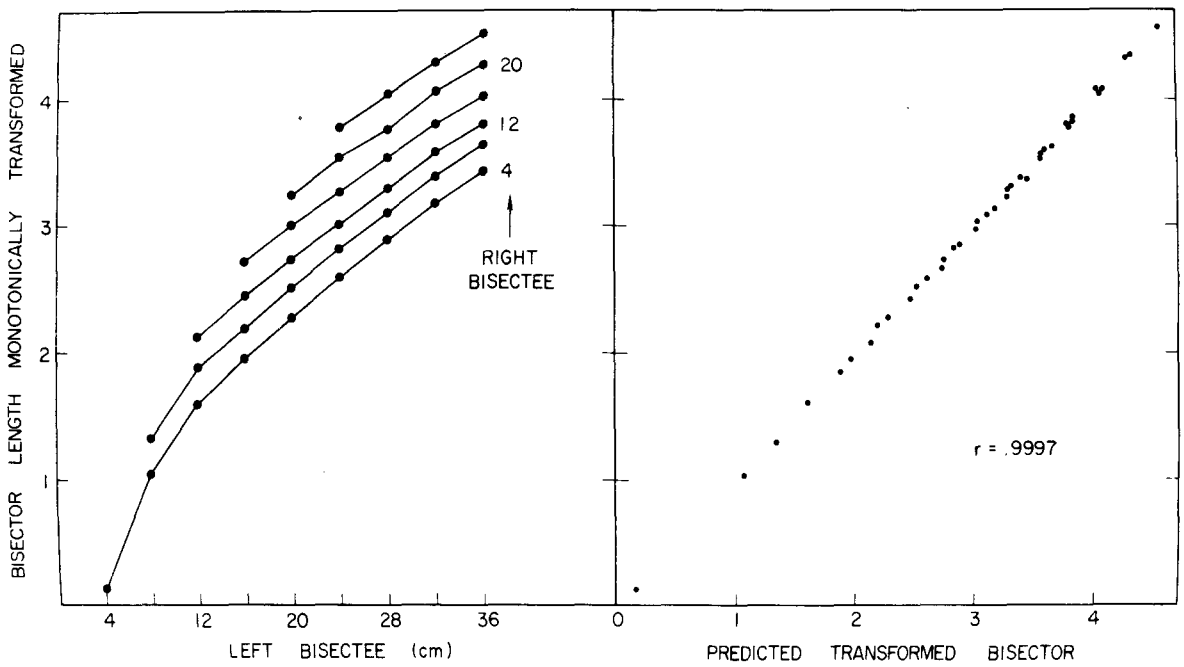

Figure 4. Left panel gives factorial-type plot of mean best monotone-additive transformation (arbitrary zero and unit) of bisector length as a function of lengths of left and right bisectees. Right panel plots same as a function of the best additive fit to the transformed data. (From Anderson, 1977.)

gurantee a high correlation. The above examples may be surprising in that the correlations are .98, .99, or even higher. They are not exceptional, however, for comparable correlations have been obtained in many other cases, even some that contained substantial crossover interactions (Anderson & Shanteau, 1977).

The fact is that the correlation coefficient is logically invalid as a general tool for testing goodness of fit. It measures the wrong thing, namely, the degree of agreement between the model and the data. A valid test must measure the degree of disagreement between the model and the data. These two measures can lead to quite different conclusions, as the above examples have shown.

This point can be formalized by noting that there is a direct relation between the predicted-observed correlation and the analysis of variance. The above procedure of fitting the additive model by least squares is equivalent to the analysis of variance model with the interaction term deleted. Thus, r^2 equals the proportion of systematic variance accounted for by the additive main effects, whereas $1 - r^2$ is the proportion of systematic variance accounted for by the interaction term. Since the interaction term measures the discrepancies from additivity, it provides the proper test of the model. Even small discrepancies can be important psychologically.

Various other statistics are functionally equivalent to the correlation coefficient. These include percentage of variance accounted for, as well as stress values in nonmetric analysis. These statistics have their uses, but they are basically misleading in model analysis, for they seem to test the model but do not really do so.

Comment on Scatterplots

The scatterplot is superior to the correlation coefficient because it presents a more detailed picture of the data. Systematic deviations from the diagonal line of perfect fit can sometimes be detected by visual inspection.

Nevertheless, scatterplots have limited value. One major shortcoming is that real discrepancies can be present that do not exhibit any systematic trend along the line of perfect fit. Visual inspection cannot distinguish such real discrepancies from error variability.

Comment on Factorial Plots

Factorial plots are superior to scatterplots because they exhibit the data as a patterned, two-dimensional function of the stimulus variables. This patterning allows easier interpretations than does the scatterplot. Further, the factorial plot presents the data themselves, whereas the scatterplot is half dependent on the model. At the same time, the prevailing irregularity in the factorial plot gives a handy visual assessment of the degree of error variability.

Surprisingly, factorial design is not overly common in psychophysics. Even when factorial design is used, its full capabilities are often neglected. In part, this may result from dominance of the one-variable approach. In the many-variable approach, as noted above, factorial design becomes a natural tool.[1]

Comment on Analysis of Variance

Various writers have expressed concern that analysis of variance does not have adequate power for model analysis. Two of the above examples

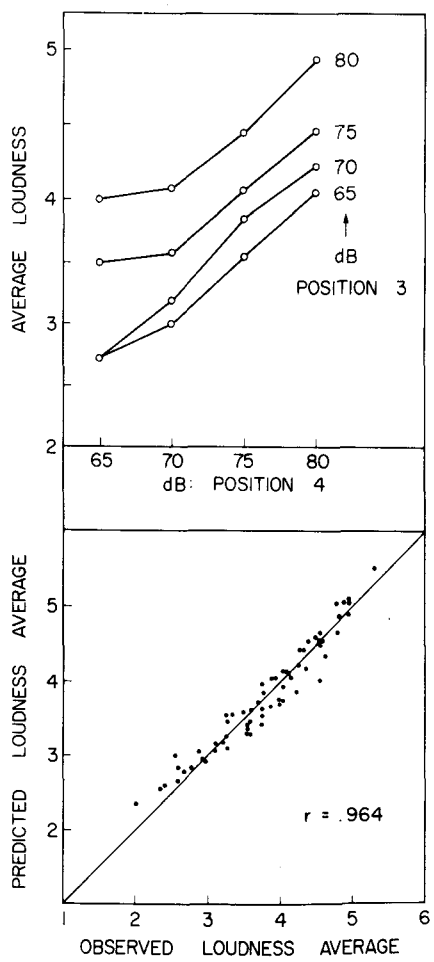

Figure 5. Upper panel gives factorial plot of mean judgment of average loudness as a function of decibel values of noises at Serial Positions 3 and 4. Lower panel plots predictions from additive model as a function of the observed means. (After Parducci, et al., 1968.)

should help alleviate this concern. In the size-weight illusion (Figure 6), the one-point discrepancy was enough to produce a reliable effect, even in the global interaction test. In length bisection (Figure 4), the very small discrepancies from additivity were readily detected even though they accounted for only .0007 of the systematic variance. It seems evident that the analysis of variance itself has ample power.

Of course, power also depends on experimental design and data reliability. Statistical methods should not be expected to compensate for weak design or for unreliable data.

PROBLEMS IN MODEL VALIDATION

Although no sharp dividing line can be drawn, there is a useful and important distinction to be made between qualitative and quantitative studies. In qualitative studies, which form the bulk of the work in any substantive area of psychology, it is often sufficient to verify a directional prediction. Thus, one experimental condition may be shown to be more effective than another without primary concern for the exact degree of difference.

Greater demands are imposed when interest shifts to the study of algebraic models or functions. Unless the model can give a fairly exact account of the data, it may mean very little.

The example of length bisection (Figure 4) provides a good illustration of the demands of model analysis. That the bisection model accounts for nearly all the systematic variance is trivial; no one could doubt that subjects can choose one length that lies roughly midway between two given lengths. The fact remains that the bisection model is invalid in this case. But this fact might have been slow to emerge without experimental design and statistical analysis of sufficient precision and cogency. Certainly, as the previous section demonstrates, a reliance on correlation-scatterplot statistics would not have done the job.

However, the correlation-scatterplot discussion touches only one aspect of the problem of validating

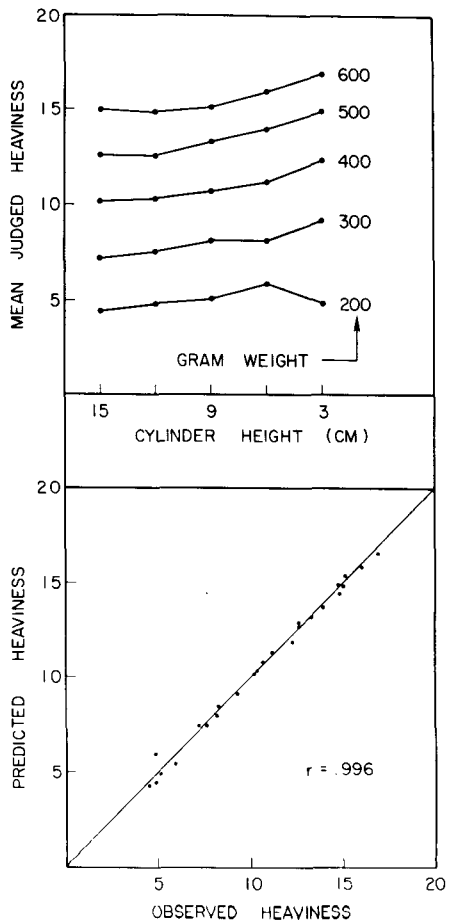

Figure 6. Upper panel gives factorial plot of mean judgments of heaviness as a function of gram weight and size of object lifted. Lower panel plots predictions from additive model as a function of the observed means in the upper panel. (After Anderson, 1972.)

algebraic models. Four related and more general aspects will be covered in this section.

The Role of Statistical Tests

The need for statistical tests of goodness of fit can hardly be overemphasized. Much current work rests on methods that, like the correlation-scatterplot statistics considered above, are fundamentally inadequate to the job. Work that relies on such methods is likely to be inconclusive, largely wasted.

However, although statistical tests are necessary, they are seldom sufficient. They suffer, indeed, from a variety of shortcomings, and some of these will be discussed here.

Accepting the Null Hypothesis

Attempts to establish algebraic integration models sometimes cause concern because accepting the model corresponds, in statistical terms, to accepting the null hypothesis in a statistical test (see Binder, 1963; Grant, 1962). In a strict sense, of course, the null hypothesis can never be proved. However, that is as true of algebraic models in physics as in psychology. In psychology, however, prevailing response variability makes statistical analysis more generally necessary.

The question of power. Unfortunately, the null hypothesis logic, which is so useful in ordinary experimental work, is awkward in testing models. In ordinary experiments, the investigator usually wishes to reject the null hypothesis. Failure to reject is ordinarily failure pure and simple, so it is advantageous to avoid weak experiments.

In model tests, however, failure to reject can seem like success. That sets up a selection tendency toward weaker experiments, or toward weaker statistical analyses. Such selection tendencies may be one contributory cause to the popularity of correlation-scatterplot analyses. Weak designs or weak analyses may be less than useless by seeming to support a bad model (Grant, 1962).

Power considerations must therefore be given primary concern in model analysis. Power depends on experimental design, and on statistical analysis in various well-known ways that need not be repeated here. However, it may be worth emphasizing that the major determinant of power is prior knowledge, the importance of which cannot be overestimated. Vague or conflicting though it often is, prior knowledge controls the choice of response measure, and details of design, procedure, and analysis. It is on these choices that power depends.

An example. A case of inappropriate acceptance of the null hypothesis by Curtis and Mullin (1975) will help illustrate some of the practical considerations that can arise. This report is relevant here because the authors were apparently trying to apply a functional measurement analysis to a loudness averaging task like that of Parducci, Thaler, and Anderson (1968). Curtis and Mullin had nine subjects use a magnitude estimation response to judge average loudness of two sounds whose decibel values were varied in a factorial design. They assumed that the averaging model with equal weighting would hold, and so the above parallelism theorem became applicable. The group analysis of variance failed to produce a significant interaction. Curtis and Mullin apparently thought that their failure to obtain a significant interaction somehow required an acceptance of the null hypothesis that the curves were truly parallel except for random variability. The parallelism theorem would then imply, in particular, that the magnitude estimation response was a linear scale.

In the functional measurement approach, contrary to Curtis and Mullin (1975), accepting the null hypothesis in this case would be most inappropriate. One major reason is that previous work with functional measurement methods has amassed extensive evidence that magnitude estimation is biased and nonlinear (e.g., Anderson, 1972, 1974a, p. 289, 1974c, p. 231, 1976b; Weiss, 1972, 1975). Real discrepancies from parallelism would therefore be expected to hold for the loudness averaging task with the magnitude estimation response used by Curtis and Mullin. Since the example of Figure 6 above shows that the analysis of variance has ample power to detect minor deviations from parallelism, it would seem that Curtis and Mullin's failure to get significant nonparallelism must be due to high variability or unreliability in their data.

This diagnosis of unreliability is supported by closer study of their results. Nonparallelism corresponds to an exponent greater than 1 in their output function. However, if a t test is performed on their reported data, the mean exponent of 1.48 is not reliably different from 1. More revealing is the 95% confidence interval; it extends from .93 to 2.03. In other words, the level of precision in their data does not define even the probable localization of the true mean exponent any better than somewhere within this interval. The extreme breadth of this confidence interval reflects both large individual differences and the high variability of magnitude estimation. Large individual differences, of course, cause low power in the group analysis of variance that Curtis and Mullin performed. For this reason, among others, it would be inappropriate to accept the null hypothesis on the basis of their analyses.

A rough rule about power. For many reasons, it is often problematical to judge how far a test that fails to reject a model allows that model to be "accepted." When a confidence interval can be set up, its width is a useful guide. A wide confidence interval would mean that the data have little bearing on the validity of the model, whereas a narrow confidence interval would support the model.

Factorial plots can also serve as a guide in the evaluation of additive or linear models. Confidence intervals may not be too useful in this case because the concern is with the overall pattern. However, the degree of irregularity in the factorial plot provides a handy visual index of power, one that is usefully supplemented by statistical tests on the interaction or components thereof.

In addition, the following rough rule about power merits consideration: Power is adequate when the discrepancies are significant statistically, but unimportant substantively. This rule may seem paradoxical, for it implies that the null hypothesis should be rejected before it can be accepted. Nevertheless, this rule seems to represent common sense and common practice. If the discrepancies are statistically significant, that is prima facie evidence for adequate power. But, if they are not substantively important, then it seems reasonable to accept the model.

Exactly this case arose in the discussion of the size-weight experiment of Figure 6 above. In a very real sense, the significant one-point discrepancy provided support for the basic integration model.

In the abstract, this rule lacks force because it lacks a criterion for importance of the discrepancies. In practice, it seems to work fairly well. In some cases, the decision is straightforward, either for rejection, as in Figures 2 and 4, or for acceptance, as in Figure 6. In other cases, as in Figure 5, decision must be deferred. And in some cases, no doubt, a discrepancy that appears unimportant today will become important tomorrow.

Overall, this rule seems to be a fair reflection of how investigators actually behave. It provides no routine recipe, but rather recognizes that research focuses on the uncertain and the partially known. There is no perfect solution to the problems of model analysis. Still, some solutions are better than others, just as the factorial plot is better than the scatterplot.

Rejecting the Null Hypothesis

If the observed data fail the test of goodness of fit, then the interpretation becomes problematical. All such tests rest on auxiliary assumptions other than the model itself. The parallelism test, in particular, rests on the three assumptions discussed above, and any one could be at fault.

The main concern, in most cases, would be that significant discrepancy signaled some basic flaw in the model itself. However, discrepancy could also result from bias in the response measure, from stimulus interaction, or from some combination of all three faults. That makes it difficult to be at all certain about the cause of the discrepancy.

No general rule seems possible, and each case needs to be considered on its individual merits. Since the five examples of Figures 2-6 all showed significant discrepancies, it is instructive to consider their interpretation.

In the grayness averaging experiment of Figure 2, the nonparallelism is marked. Weiss' interpretation was that the averaging model itself was basically correct, and that the discrepancy reflected bias in the magnitude estimation response. This interpretation was supported by the success of the model when a rating response was used.

In the grayness bisection experiment of Figure 3, the nonparallelism is not large and might not seem too serious. However, it makes a substantial change in the psychophysical law. In this experiment, it was possible to transform the data to parallelism using a monotone transformation (see below). The test of fit showed nonsignificant discrepancies, in support of the averaging model for bisection. The functional scale from the bisection task yielded an exponent of .2 in a power function fit. This contrasts with the exponent of .33 from the Munsell scale, and the exponent of 1.2 obtained from magnitude estimation (Stevens, 1974). Cross-task validation was provided by the scale agreement across bisection, averaging, and differencing tasks (Anderson, 1976b). In this case, therefore, the bisection model seems to be correct, whereas the Munsell scale shows a moderate bias. In effect, the steps on the Munsell scale become progressively too small toward the white end of the scale.

In the length bisection data of Figure 4, the deviations from parallelism are small. However, it should be recognized that these data are as additive as a monotonic transformation can make them. That significant nonadditivity remains argues strongly against the bisection model. It also argues for the power of even relatively small factorial-type designs to resist monotone transformation to additivity.

In the loudness averaging experiment of Figure 5, the interpretation remains uncertain. Attempts to transform the response in the original report were unsuccessful, but the transformation method was not optimal, so the fault may yet lie in the response scale. The loudness averaging study of Curtis and Mullin (1975) failed to note the discrepant results in Parducci et al. (1968), and simply assumed the averaging model to be true. That is of course in line with the success of the model on other psychophysical dimensions, such as grayness, heaviness, and length. However, some caution must be used in generalizing these results to loudness averaging, because it is also possible that differential weighting is involved, with louder sounds receiving greater weight. Differential weighting would cause the convergence interaction visible in the upper panel of Figure 5. This possibility requires serious consideration, because differential weighting has been fairly frequent, at least with verbal and symbolic stimuli (Anderson, 1974d). If differential weighting is involved, then it would be wrong to transform the data to additivity. The dilemma posed by the Parducci et al. (1968) experiment thus remains unresolved.

In the size-weight data of Figure 6, the significant discrepancy could be localized in a single point. Further, this discrepant point could be given a reasonable interpretation in terms of an end-effect in the response scale. In this case, therefore, it seemed reasonable to discount the significant discrepancy.

Discounting significant discrepancies may seem cavalier, but it makes sense. The model itself is considered to apply to a pure or idealized case, just as in physics. Thus, the law of falling bodies, or the law of the pendulum, refer to idealized cases in which friction is absent. In psychology, similarly, minor biases in the response scale, such as number preferences, are inevitable and they will produce significant discrepancies if sufficient data are collected. Similarly, the independence assumption of no stimulus interaction will probably never be perfectly true. It is on this reasoning that the above rough rule about power is based.

Error Theory for Nonmetric Analysis

Nonmetric analysis, as is well known, has been severely handicapped by lack of ways to handle response variability. Conjoint measurement and most multidimensional scaling methods rely on nonmetric analysis of certain algebraic models. Without goodness of fit tests, the validity of these models remains in doubt. The existence theorems of conjoint measurement can be seen essentially as goodness of fit tests that rely only on ordinal properties of the data. However, the theorems assume errorless data and are difficult to use with real data.

Recent work in functional measurement has suggested ways by which a valid and general error theory for nonmetric data might be obtainable. This approach is based on joint application of functional measurement and multidimensional scaling. Two methods will be discussed briefly here.

The first method was used in the experiment on grayness bisection of Figure 3. Each subject served in 10 replications of the basic 6 by 5 design so the analysis could be carried out at the individual level. If the averaging model holds for bisection, and if the response measure is a linear scale, then the data will exhibit parallelism. The physical measure, namely, the reflectance of the gray chip chosen as the bisector, is of course not a linear function of grayness and did not exhibit parallelism. However, the Munsell scale also failed the test of parallelism. This leaves the interpretation uncertain: The basic model might be wrong, or the Munsell scale might be nonlinear. A more definite analysis can be obtained as follows.

The first step is to apply a monotone transformation to make the data as additive as possible. This can be accomplished with one of the transformation programs developed in the work on multidimensional scaling, in this case the ADDALS program of de Leeuw et al. (1976). If grayness bisection does follow the averaging model, then the transformed data should exhibit parallelism except for unsystematic error variability.

The next step, therefore, is to assess the deviations from parallelism. This presents a problem. The monotone transformation is, in effect, a curve-fitting procedure that has a large, unknown number of parameters. It has great power to force parallelism onto the data. If systematic discrepancies remain, they will generally be small and difficult to assess statistically.

Further, the analysis of variance cannot be applied in the usual manner. In the first place, the monotone transformation has used up an unknown number of degrees of freedom. In addition, the transformed data are no longer independent, being intercorrelated through their dependence on a common transformation.

Fortunately, there is a simple way around these problems. The monotone transformation program is applied separately to each replication of the design. Since the transformed data are then independent across replications, any one degree of freedom component of the interaction has a valid test. For example, the algebraic value of the Linear by Linear component could be calculated for each replication. The model-implied null hypothesis, that the true mean of these algebraic values is zero, can readily be tested, by t ratio, say, or by nonparametric test.

More generally, the overall Row by Column interaction should be nonsignificant when tested against its proper error term, namely, the Row by Column by Replication interaction. The entries within each replication are all intercorrelated, of course, but the repeated measurements analysis of variance allows for that. This same approach may also be used for group analyses by treating subjects in the same way as replications.

The second method relies on two-stage integration models (Anderson, 1974a, pp. 251-258). These refer to tasks that incorporate two integration operations. Applied to the present problem, the essential idea would be to use one operation as the frame for transforming the response, the other operation as the base for testing goodness of fit.

In the three-factor adding model, A + B + C, for example, the response would be transformed to make A + B maximally additive. If the model is correct, then this transformation is generally possible and the transformed response will ideally be a linear scale. Accordingly, the interaction between C and the compound variable (A + B) should be nonsignificant. If either operation is nonadditive, then this test should fail. Psychophysical averaging is well suited to this method because three or more stimulus factors can readily be employed, even with, indeed, especially

with, different stimulus dimensions (Anderson, 1974c).[2]

A variant of this second method is possible when the two operations represent distinct tasks with the same stimuli. In some cases, the functional stimulus values should be the same in both tasks. In the grayness bisection experiment of Figure 3, for example, subjects also rated average grayness and difference in grayness between each pair of chips. All three tasks yielded the same grayness scale (see Figure 7 below), which supports the application of monotone transformation to the bisection data. This approach has also been explored by Birnbaum and Veit (1974).

These methods have not been adequately explored in practice, and their use requires caution. One main concern is over possible bias in the monotone transformation which could arise, for example, from discreteness in the design. This problem has not received adequate study in the literature on nonmetric multidimensional scaling which has been oriented more toward the discovery of dimensions than to the algebraic models themselves which are the main concern of functional measurement.[3,4]

A second main concern is that monotone transformation may be too effective in the sense that it eliminates real discrepancies (Anderson, 1962b, p. 410). Information available on this problem is hopeful, but largely tangential. Work on nonmetric multidimensional scaling has shown that ordinal information provides strong constraints (e.g., Shepard, 1962, 1966). However, most of this work has been concerned with extraction of dimensions, and that is much less demanding than reconstruction of the exact metrics.

Furthermore, illustrations that nonmetric analysis can reconstruct additive metric data merely from rank orders (e.g., Weiss & Anderson, 1972) are not squarely to the point. What needs to be studied is how strongly data that are inherently nonadditive resist being transformed to additivity, and what design considerations are necessary to avoid inappropriate transformation.

The two bisection examples (Figures 3 and 4) have twofold relevance here. First, they show that simple additive models are sometimes, but not always, correct, even with a fixed judgment task. Second, the length bisection data did successfully resist transformation to additivity. Even a relatively small design may be adequate, therefore, at least with data that are as highly reliable as bisection data.

Nevertheless, much more work is needed on this question. On the basis of present knowledge, failure to find significant nonadditivity in the transformed data may have rather little bearing on the question of whether the process is truly additive.

The danger of inappropriate transformation can be reduced by transforming more than one replication at a time, by increasing the number of levels in each factor of the design, and, perhaps most effectively, by increasing the number of factors in the design. However, studies with artificial data are greatly needed to assess minimum design requirements that will allow real deviations from the model to be detected. These studies need to be done for models that are empirically reasonable, and inherently nonadditive, such as the averaging model with unequal weighting, and the ratio model developed in information integration theory (Anderson & Farkas, 1975; Leon & Anderson, 1974; Oden, 1974).[5]

Nonmetric Smoothing and Parameter Estimation

Integration models may be studied for their own sake, or they may be used as the base and frame for scaling. In the size-weight illusion, for example, main interest centers on how the visual and kinesthetic cues are integrated. Psychophysical averaging tasks, on the other hand, have less intrinsic interest and are usually employed as tools, for cross-task validation, for example, or for scaling.

Scaling and model testing involve somewhat different considerations that affect both experimental design and data analysis. Different designs and different statistical analyses may be indicated, depending on which goal is primary. One problem of scaling has special relevance to the present discussion.

Scaling is essentially a matter of parameter estimation, at least from the standpoint of data analysis. Parameter estimation is an intricate subject, much studied in statistical theory, but lucid reviews are given by Bush (1963) and Restle (1971). As is well known, the statistical properties of the estimates from a given set of data will depend on the method employed in the estimation. The reliability of the estimates, in particular, will be greater or lesser, depending on the estimation method. Only the problem of monotone smoothing will be considered here.

If the linear model is correct, and if the response measure is on a linear scale, then the marginal means of the factorial design are unbiased estimates of the stimulus values as noted above. Of course, these estimates still contain variability and to that extent are unreliable. Smoothing the data can increase reliability.

Smoothing for the linear model is done by applying a transformation that makes the data more additive or parallel. Even when the integration process is additive, the data themselves will not be exactly parallel because of the prevailing response variability. The transformation reduces this variability, making the data more parallel, and so also reduces the unreliability of the parameter estimates.

There is nothing mysterious about data smoothing. Essentially, it is a matter of fairing a curve through variable data points. The only difference in the present case is that the "curve" is actually the integration function.

Nonmetric methods have special usefulness for data smoothing, in part because of the availability of powerful computerized techniques, in part because such transformations have optimal properties. An illustrative application is given in Weiss and Anderson (1972). The original metric data satisfied the parallelism test. These metric data were reduced to rank orders, and Kruskal's (1965) MONANOVA was applied. Even in this rather small 5 by 5 design, it was possible to reconstruct the metric data from the rank orders. The reconstructed data, of course, were even more parallel than the original data (Weiss & Anderson, 1972, Figures 1 and 2).

Nonmetric smoothing has an added advantage. Not only will it reduce variability, but it will also reduce bias. Any numerical response will have some bias, from number preferences, for example, even though the bias may not be large enough to reach statistical significance. Monotone transformation to additivity will reduce this bias and so improve the estimates.

Of course, nonmetric smoothing depends completely on the integration function. This point cannot be overemphasized. To the degree that the model is wrong, smoothing will tend to inject an additional bias of its own. Accordingly, an adequate test of goodness of fit is essential. This represents a serious problem because, as noted above, monotone transformation may too readily impose additivity where it does not hold.

THE PROBLEM OF GENERALITY

Within any one experimental task, the test of goodness of fit can be viewed as a test of internal structure or consistency. It asks whether the given data satisfy the pattern or structure implied by the model. For additive models, this pattern is one of parallelism, as noted above. Other models imply other patterns (Anderson, 1974a, p. 264, Note 2; Shanteau & Anderson, 1972). Such tests of internal structure are the necessary first step in the study of algebraic models.

However, external consistency is also needed. No one experiment goes very far by itself. Only an interlocking body of experiments can provide an adequate theoretical base (see, e.g., Anderson, 1962b, 1974a; Birnbaum & Veit, 1974; Cliff, 1973, p. 480; Garner, Hake, & Eriksen, 1956, among others).

Three kinds of external consistency have been of concern in the functional measurement approach. Each deserves a brief remark.

Cross-Task Stimulus Consistency

The first kind of external consistency concerns the invariance of the subjective stimulus scale across different tasks. Results so far have been reasonably promising for psychophysical stimuli.[6]

Work on heaviness (Anderson, 1971, 1972) has obtained an interesting cross-task consistency. The same heaviness scale was obtained from the size-weight illusion as from an averaging task. This is important because the integration is preconscious in the former case, postconscious in the latter case. Accordingly, the agreement of these two scales would seem to be an important clue to the processing structure. Similar studies in other stimulus domains, taste and warmth, for example, would be desirable.

The work on grayness (Anderson, 1976b; Weiss, 1972, 1975) has special interest because the bisection task yielded the same grayness scale as the differencing and averaging tasks. This can be seen in Figure 7, which plots the psychophysical law obtained from the three tasks: The three curves are virtually identical.

This cross-task consistency is important because the perceptual-cognitive demands of the three tasks are quite different. Bisection asks for a physical response that makes equal two direct sensory differences. Differencing asks for a verbal response to represent the magnitude of one sensory difference. Averaging also uses a verbal response but a rather different cognitive operation. That the same grayness scale is operative in all three tasks is of course not surprising. However, the capability of functional measurement to establish this cross-task consistency is notable.

This grayness scale has an exponent of about .2 in a power function fit. That contrasts moderately with the exponent of .33 for the Munsell scale, and sharply with the exponent of 1.2 obtained from magnitude estimation. Since the present determination of the psychophysical law satisfies both within-task consistency and between-task consistency, it seems reasonable to think that it is correct.

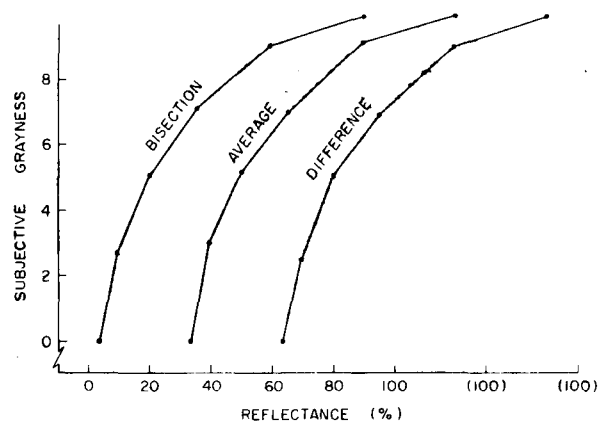

Figure 7. The psychophysical law for grayness from bisection, averaging, and differencing tasks. Curves displaced horizontally. (From Anderson, 1976b.)

Cross-Task Response Consistency

The second kind of external consistency concerns the generality of the rating method. According to the functional measurement logic outlined above, success of the parallelism test for the additive model supports the observed response measure as a valid, linear scale. The same logic applies to corresponding tests for the various nonadditive integration models.

Following this logic, substantial evidence has accumulated to validate ratings as true linear scales in numerous different tasks, not only in psychophysics (Anderson, 1974a), but also in many other areas (Anderson, 1974b, d; 1976a). This support for the rating method is both broad and firm.

Of course, as noted above, certain modest experimental precautions are indicated in order to avoid the various known biases that can contaminate the ratings. The development of such precautions is an important methodological contribution of the program of research on information integration theory. This methodological task should not be considered complete, but a matter for continuing study. It is important because the ability to rely on numerical response methods yields far more rapid progress than is possible with nonmetric analysis. Further, the development of numerical response methods is important, if indeed not vital, for the analysis of stimulus interaction.

Model Generality

The third type of external consistency concerns the generality of the integration function. Workers in many separate areas have conjectured that simple algebraic models might be operative in perception and judgment (see, e.g., Anderson, 1970, 1974a, 1975, p. 480). However, the analysis of these models was held back by the lack of adequate measurement methodology. Although many investigators have espoused ideas similar to those used in functional measurement, the dominating orientation saw measurement as a methodological or mathematical preliminary to substantive inquiry.[7] In contrast, the functional measurement approach makes measurement an integral, organic part of the ongoing empirical investigation.[8]

With functional measurement methods, the study of algebraic models can be placed on a rigorous, quantitative basis. Numerous experimental studies have given strong support to such models across many different areas, including psychophysics, psycholinguistics, decision theory, and social perception (Anderson, 1974a, b, d, Note 3). Thus, there does seem to be substantial model generality. Overall, the accumulated evidence points to the existence of a general cognitive algebra.

REFERENCE NOTES

1. Anderson, N. H. *Methods for studying information integration* (Technical Report CHIP 43). La Jolla, Calif: Center for Human Information Processing, University of California, San Diego, June 1974.

2. Anderson, N. H. *Algebraic models for information integration* (Technical Report CHIP 45). La Jolla, Calif: Center for Human Information Processing, University of California, San Diego, June 1974.

3. Anderson, N. H. *Social perception and cognition* (Technical Report CHIP 62). La Jolla, Calif: Center for Human Information Processing, University of California, San Diego, July 1976.

REFERENCES

ANDERSON, N. H. Applications of an additive model to impression formation. *Science*, 1962, **138**, 817-818. (a)

ANDERSON, N. H. On the quantification of Miller's conflict theory. *Psychological Review*, 1962, **69**, 400-414. (b)

ANDERSON, N. H. Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, 1965, **70**, 394-400.

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

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

ANDERSON, N. H. Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press, 1974. (a)

ANDERSON, N. H. Cognitive algebra. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 7). New York: Academic Press, 1974. (b)

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

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

ANDERSON, N. H. On the role of context effects in psychophysical judgment. *Psychological Review*, 1975, **82**, 462-482.

ANDERSON, N. H. How functional measurement can yield validated interval scales of mental quantities. *Journal of Applied Psychology*, 1976, **61**, 677-692. (a)

ANDERSON, N. H. Integration theory, functional measurement, and the psychophysical law. In H.-G. Geissler & Yu. M. Zabrodin (Eds.), *Advances in psychophysics*. Berlin: VEB Deutscher Verlag, 1976. (b)

ANDERSON, N. H. Failure of additivity in bisection of length. *Perception & Psychophysics*, 1977, in press.

ANDERSON, N. H., & FARKAS, A. J. Integration theory applied to models of inequity. *Personality and Social Psychology Bulletin*, 1975, **1**, 588-591.

ANDERSON, N. H., & SHANTEAU, J. Weak inference with linear models. *Psychological Bulletin*, 1977, in press.

BINDER, A. Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 1963, **70**, 107-115.

BIRNBAUM, M. H., & VEIT, C. T. Psychophysical measurement: Information integration with difference, ratio, and averaging tasks. *Perception & Psychophysics*, 1974, **15**, 7-15.

BOGARTZ, R. S., & WACKWITZ, J. H. Polynomial response scaling and functional measurement. *Journal of Mathematical Psychology*, 1971, **8**, 418-443.

BUSH, R. R. Estimation and evaluation. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: Wiley, 1963.

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

CURTIS, D. W., & MULLIN, L. C. Judgments of average magnitude: Analyses in terms of the functional measurement and two-stage models. *Perception & Psychophysics*, 1975, **18**, 299-308.

DE LEEUW, J., YOUNG, F. W., & TAKANE, Y. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 1976, **41**, 471-503.

DURLACH, N. I., & COLBURN, H. S. Binaural phenomena. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 4). New York: Academic Press, in press.

GARNER, W. R., HAKE, H. W., & ERIKSEN, C. W. Operationism and the concept of perception. *Psychological Review*, 1956, **63**, 149-159.

GRAESSER, C. C., & ANDERSON, N. H. Cognitive algebra of the equation: Giftsize = generosity x income. *Journal of Experimental Psychology*, 1974, **103**, 692-699.

GRANT, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 1962, **69**, 54-61.

KRANTZ, D. H. Measurement structures and psychological laws. *Science*, 1972, **175**, 1427-1435.

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.

LEON, M., & ANDERSON, N. H. A ratio rule from integration theory applied to inference judgments. *Journal of Experimental Psychology*, 1974, **102**, 27-36.

MARKS, L. E. *Sensory processes*. New York: Academic Press, 1974.

NORMAN, K. L. A solution for weights and scale values in functional measurement. *Psychological Review*, 1976, **83**, 80-84.

ODEN, G. C. *Semantic constraints and ambiguity resolution*. Unpublished doctoral dissertation, University of California, San Diego, 1974.

PARDUCCI, A., THALER, L., & ANDERSON, N. H. Stimulus averaging and the context for judgment. *Perception & Psychophysics*, 1968, **3**, 145-150.

RESTLE, F. *Mathematical models in psychology: An introduction*. Baltimore: Penguin Books, 1971.

SHANTEAU, J. Component processes in risky decision making. *Journal of Experimental Psychology*, 1974, **103**, 680-691.

SHANTEAU, J. C., & ANDERSON, N. H. Integration theory applied to judgments of the value of information. *Journal of Experimental Psychology*, 1972, **92**, 266-275.

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962, **27**, 219-246.

SHEPARD, R. N. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 1966, **3**, 287-315.

STERNBERG, S. The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, 1969, **30**, 276-315.

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

STEVENS, S. S. Perceptual magnitude and its measurement. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press, 1974.

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

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

WEISS, D. J. FUNPOT: A FORTRAN program for finding a polynomial transformation to reduce any sources of variance in a factorial design. *Behavioral Science*, 1973, **18**, 150.

WEISS, D. J. Quantifying private events: A functional measurement analysis of equisection. *Perception & Psychophysics*, 1975, **17**, 351-357.

WEISS, D. J., & ANDERSON, N. H. Use of rank order data in functional measurement. *Psychological Bulletin*, 1972, **78**, 64-69.

NOTES

1. An interesting illustration of the potential of factorial design for substantive analysis arose in conversations with Edward Carterette about perceptual "trading relations." In binaural localization (Durlach & Colburn, in press), for example, the loudness difference and the time difference between the sounds to the two ears act as joint cues; their integrated resultant determines the phenomenal localization. The trading relation approach studies what magnitude of one cue is needed to just offset a given magnitude of the other cue. This information is not adequate to determine the cue integration function because, in particular, it does not look at the effect of both cues acting in the same direction. Neither is it adequate to determine the subjective values of the separate cues.

In the integration-theoretical approach, the two cues would be varied independently in a factorial-type design, and the response would be the localization azimuth. It seems reasonable to expect an averaging rule for cue integration. Factorial-type design would allow a test of this hypothesis which, if successful, would also yield the psychophysical functions for the separate cues. The trading relation, if desired, could be derived from this information, for it is merely the plot of one psychophysical function against the other.

2. A unique possibility for cross-stimulus comparison may be obtainable from the averaging model. With suitable design, the functional scales of loudness, brightness, sweetness, pitch, etc., are on interval scales with common zero and common unit. One specific experimental approach would use a heteromodal averaging task (Anderson, 1974c) in which subjects would judge average magnitude of sets of stimuli from several different dimensions. Variations in set size can produce the differential weighting and design constraints that are needed to obtain uniqueness (Anderson, 1974a, p. 227). Some estimation problems are discussed by Norman (1976). If the cross-modal averaging model holds empirically, it can provide direct comparisons of magnitude between quite different stimulus dimensions.

3. These two approaches, both of which make use of algebraic models and monotone response transformation, were introduced independently in the same year (Anderson, 1962a, b; Shepard, 1962) and have remained largely independent ever since. This reflects their difference in orientation, toward dimensional stimulus representation, on the one hand, and toward stimulus integration, on the other.

This difference in orientation can be highlighted by the corresponding treatment of discrepancies. In multidimensional scaling, discrepancies are usually (though not necessarily) taken as evidence for additional dimensions. Adding another dimension reduces the discrepancies but does not change the basic model. In integration theory, however, significant discrepancies that remain after monotone response transformation would usually be taken as evidence against the model itself. The study of length bisection (Figure 4) is a good example.

4. Sternberg's (1969) important additive-factor method for reaction time uses interactions from factorial design as clues to interactions among sequential processing stages. By nature of the focus on the additivity of processing times across successive stages, response transformation is not allowable, a property of Sternberg's method that avoids one difficulty in the interpretation of the factorial interactions. On the other hand, the method does not apply to general problems of stimulus integration (see also Anderson, 1974a, pp. 271-274).

5. Although not immediately relevant to the discussion of the text, it may be appropriate to add here a note on the statistical analysis of the linear fan prediction of the multiplying model. For individual subject analyses, there is usually no problem, since the within-cell variability may usually be used to test both the Linear by Linear and the residual components of the interaction. For group data in repeated measurements designs, a valid test of the Linear by Linear component may be obtained by calculating the algebraic value of this component for each subject and testing the

null hypothesis that the mean of these algebraic values is zero. However, the test of the residual is biased. Previously, it had been thought that this bias was negligible (Graesser & Anderson, 1974, appendix), but that conclusion was based on an error, and later applications have shown that the bias is not always negligible. The present status of the omnibus test on the residual is thus uncertain. Valid tests of the residual can be obtained by extraction of Linear by Quadratic and other higher order components of the interaction, though perhaps with some loss of power.

6. One particular kind of cross-task stimulus consistency deserves comment. Judgments are often obtained, not only of the stimulus combinations, but also of the separate single stimuli. These single stimulus judgments are often treated as scale values, to be used in testing the integration rule. Surprisingly, this direct approach is inferior to the indirect approach in which the stimulus values are estimated from the judgments of the combinations (Anderson, Note 1). Among other reasons, the variability of the single-stimulus judgments reduces power, at least as the test is usually performed.

At the same time, judgments of the single stimuli and of the combinations can be seen as two different tasks, for the latter involves an integration operation and the former does not. Thus, it is a proper question whether the stimulus scale is the same in both tasks. Some results from decision theory are given by Shanteau (1974), but there is little firm evidence on this question in psychophysical judgment.

7. This point is well illustrated in Stevens' (1971, p. 431, 1974) reaction to algebraic models as the basis for measurement. Finding that magnitude estimation did not satisfy these models, Stevens concluded that the judgment task introduced a bias. This interpretation hardly seems tenable in view of the extensive success of rating methods in these tasks.

8. The issues of response and model generality also bear on the problem of monotonic indeterminacy (see Anderson, 1974a, p. 231), which is illustrated by the classical question of whether subjects instructed to judge ratios are really judging ratios or differences (Birnbaum & Veit, 1974; Torgerson, 1961). A similar problem arises in averaging theory, since the data of Figure 2 above could be interpreted to support magnitude estimation jointly with the geometric mean which corresponds to a multiplicative model (Weiss, 1972). If monotone transformation is allowed, additive and multiplicative models may be equivalent; parallel curves that support the arithmetic mean are monotonically equivalent to a linear fan of curves that would support the multiplicative geometric mean.

But the geometric mean is not defined for negative values such as are obtained in many integration tasks, most notably in person perception. The success of the rating response and the arithmetic mean in these tasks where the geometric mean cannot apply may thus resolve the question for other tasks where both can apply. Attending to a broader range of tasks, therefore, allows progress on some of the problems of monotonic indeterminacy (Anderson, 1974a, p. 231).

(Received for publication January 30, 1976;
revision received December 28, 1976.)