JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 1976, 25, 123-128 NUMBER 1 (JANUARY)

THE SCHEMAPIRIC VIEW. NOTES ON S. S. STEVENS' PHILOSOPHY AND PSYCHOPHYSICS[1]

PETER KILLEEN[2]

ARIZONA STATE UNIVERSITY

S. S. Stevens will be remembered as the champion of operationism in psychology, as the inventor of the theory of scale types, and as the father of the psychophysical law bearing his name (Miller, 1974). Both operationism and the theory of scale types were meta-scientific enquiries into the relations between formal systems and data. It is Stevens' Power Law, and the psychophysical investigations involved in its formulation and verification, that remain a monument to schemapiric science at its best. The term schemapiric is Stevens' (1968), and it denotes the marriage of formal systems and models—schemata— to observations and tabulations—empirics. This union of the ideal and the real is found whenever there is progress in science; it characterizes the work of Stevens' later years, presented simply and clearly in the posthumous *Psychophysics.*

There is much about Stevens of which experimental analysts of behavior would approve. He was a behaviorist: "We study the response of an organism, not some nonphysical mental stuff that by definition defies objective test" (p. 51). Not a radical behaviorist to be sure, but a man with whom we could feel philosophically at ease, if not somewhat jealous of, for his simpler "conventional" behaviorism. He was an exceptional psychophysicist, more interested in substantive issues than in methodology (in one passage he criticized signal detectability theorists for "much honing of the tool's edge, but little cutting" (p. 178). Compare Day's characterization of behavior analysts as pragmatists: "The radical behaviorist [is convinced] that if knowledge is to be trusted it is often likely to lead to effective action" (1969, p. 318). Stevens approved of small N research: "Large numbers do not protect against systematic errors" (p. 293), and, through the mask of "an extreme stochastophobe", he queried "What scientific discoveries owe their existence to the techniques of statistical analysis or inference?" If few, whence the charm of statistics? "For some stochastophiles that appeal may have no deeper root than a preference for the prudent posture at a desk as opposed to the harsher, more venturesome stance in the field or the laboratory" (1968, p. 853; cf. Skinner, 1958). Stevens rearranged his data visually to make them optimal discriminative stimuli for his verbal behavior, and in the process reputedly used almost as much logarithmic graph paper as Skinner did cumulative records.

Stevens professed the importance of schemata, and it is here that some behaviorists will flinch. Data collection unguided by theory was as fruitless, in Stevens' eyes, as theorizing unguided by data: "Numbers gathered without some knowledge of the regularity to be expected almost never speak for themselves" (Kuhn, cited in Stevens, 1968). Perhaps Platt said it best: "We speak piously of taking measurements and doing small studies that will 'add another brick to the temple of science.' Most such bricks just lie around the brickyard (1964, p. 351; cf. Forscher, 1963)." For these men, effective science starts with specific questions; these lead, by a process of inference, to relevant experimentation, which in turn provokes additional questions and extends the purview of inference. Deviation too far in either direction—theoretical or empirical—breeds empty verbiage or pointless experiments. In citing von Neumann's view of the evolution of mathematics, Stevens reminds us of the Skinnerian's primary interest in verbal

[1]Stevens, S. S. *Psychophysics: introduction to its perceptual, neural, and social prospects.* Geraldine Stevens, Ed. New York: Wiley-Interscience, 1975. Pp. iv + 330, $19.95.

[2]Reprints may be obtained from the author, Department of Psychology, Arizona State University, Tempe, Arizona 85281.

behavior that is controlled by directly observed events (tacts) rather than by other words (intraverbals, echoics, and textuals): "As a mathematical discipline travels far from its empirical source, there is grave danger that the subject will develop along the line of least resistance, that the stream, so far from its empirical source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganized mass of details and complexities" (von Neumann, cited in Stevens, 1968). Schemata, be they mathematical models or verbal descriptions, are most trustworthy when most closely related to data.

But it is not only formal systems that become disengaged from their basis in empirical issues; empirical research can also become autonomous, with scores of experiments often devoted to ramifications of a provisional, and perhaps obsolete, hypothesis. There have, for instance, been many interesting experiments on conditioned suppression and facilitation, but few that address the issue that gave the paradigm birth: is it a good way to operationalize anxiety? Many experiments have employed concurrent-chain schedules, but seldom raised is the issue of their effectiveness as a measure of choice. And so on. Throughout, tactics of research are refined while strategy is ignored. In an exceptional schemapiric analysis of schedule effects, Jenkins observed that "Instead of checking, revising, and adding to the principles put forth by Skinner in 1938, many have been satisfied to generate behavioral regularities by the use of experimental arrangements far too complex to analyze" (1970 p. 106). Because of his enduring interest in the interface between data and theory, and his scepticism when discussion strayed too far from that boundary, memory of Stevens and his schemapiric view may save us from excesses of both the right and the left.

It is in the details of Stevens' own work, not in his philosophy of science, that we may take the greatest pleasure. He asked basic questions, he provided simple answers, and those are the hardest things to do. In 1953, Stevens originated a technique for sensory scaling that he called "magnitude estimation". Subjects were simply asked to estimate with numbers how loud, bright, painful, odorous, long, red, cold, heavy, viscous, rough, sweet, or fast various stimuli were. Could such a simple procedure work? Stevens presents a list of historical objections to the attempt to measure sensation, from James in 1890 to Savage in 1970. The list serves him as a foil, for he has, by this point in the book, presented many orderly graphs of data resulting from direct sensory scaling. For most perceptual continua, the graphs display power functions relating magnitude of the stimulus to magnitude of the sensation. The use of numbers as the response is not an Achilles heel for his theory, for similar functions are obtained when people are instructed to "squeeze this handgrip as hard as that light is bright" (cross-modal matching). Indeed, the exponents for the cross-modal matches may be predicted directly from the exponents for magnitude estimation, once the exponent for the continuum used to register a response is taken into account.

But reliability and consistency do not add up to validity. Is it really sensation that Stevens is measuring? Shepard (1966) noted: "Unfortunately, the evidence on this question [the form of the psychophysical function] is indirect, at best, since the internal psychological variable, Ø, is of course never itself observed, but only inferred from the various overt physical responses made by the subject. Nevertheless, as Stevens has pointed out, it may here (as elsewhere in science) be useful to introduce such a hypothetical, intrinsically unobservable variable if such a step sufficiently simplifies the relations among observable variables." Shepard goes on to demonstrate that weaker assumptions about internal events could provide an adequate basis for Stevens' power functions; sensations may be related to magnitude estimates of them by no more than monotonic transformations, and under many conditions power functions might still be expected (*cf.* Krantz, 1972).

The introduction of hypothetical constructs has some precedent in the experimental analysis of behavior: " 'Drive' is a hypothetical state interpolated between operation and behavior and is not actually required in a descriptive system. The concept is useful, however, as a device for expressing the complex relation that obtains between various similarly effective operations and a group of co-varying forms of behavior" (Skinner, 1938, p. 368). *Reification* of such constructs has traditionally been anathema, leading some to view the use

of any new construct with suspicion; but it must be remembered that "operant" is no less of a construct than "image", and demands no less care in its use (*cf.* MacCorquodale and Meehl, 1948). The ultimate validity of either construct is our ultimate—most distant—concern. Establishing useful constructs—consistent, reliable ways of grouping data whose description requires less information than do the original data—is of more immediate importance, and provides the only solid foundation upon which the construct of validity may eventually rest.

It is in any case easy to accept Stevens' data and techniques without embracing mentalism. Zuriff (1972) suggested that we think of the behavior of estimating magnitude as an example of transfer of training; just as we estimate the length of a line, we may estimate the loudness of a tone, with the behavior shaped by the use of yardsticks in the former case generalizing (a "metaphorical extension") to the intensity of a tone in the latter. The assumption of an internal, psychological scale that mediates the responses then becomes supererogatory, just as it essentially did in Shepard's analysis.

None of this would bother Stevens, whose lifelong quarry was the discovery of regularities—"invariances"—in data structures, and who took the operation of matching as the logical basis of measurement, including measurement of sensation: "Magnitude measures derive from direct cross-modality matching of magnitude or degree on one continuum to the same aspect on another. One of the continua may be the number continuum" (p. 230). Such generalizations did not come easy, however, for there were many impediments to a unified picture of psychophysical behavior. Three common scaling techniques—Thurstone scales, category (rating) scales, and magnitude estimation scales—all yielded different "psychophysical laws" for some continua, and similar "laws" for other continua. The latter are in the minority, and are exemplified by auditory pitch, apparent position, and apparent inclination. Stevens called these metathetic continua, and characterized them as "qualitative" in nature. Continua in the remaining class, which includes brightness and loudness, he labelled prothetic, and characterized as being "intensive" in nature. On prothetic continua, both category scales and Thurstone scales are curved—concave down—when plotted against a magnitude estimation scale. Thurstone scales are constructed by assuming that the psychological size of the just-noticeable-difference (JND) is constant, so that by adding up JNDs, we can find the measure of sensation. This was Fechner's insight, and was extended by Thurstone to "continua", such as excellence of handwriting, for which there are no obvious physical measures. Thurstone's contribution was the employment of the standard deviation of pair-comparison judgements as the unit for his scale, assuming that "equally often noticed differences are equal".

Are Thurstone scales more or less accurate than magnitude scales? Gösta Ekman examined a more tractable issue: are all JNDs equal in subjective size? Ekman (1956, 1959) found this basic assumption to be incorrect, but noted an important regularity in his data that could take its place: the ratio of the JND to stimulus magnitude is constant, when both are measured in psychological units. This "psychological Weber Law" is even more powerful than the original Weber Law, which stated that: the ratio of the JND magnitude to the stimulus magnitude is constant, when both are measured in physical units. More powerful, because the Weber fraction differs from one continuum to another, but Ekman's fraction is constant over at least nine continua (Teghtsoonian, 1971). The psychological size of a JND is always 3% of the psychological magnitude of the stimulus about which it is centered. *There* is a striking and beautiful invariance, one of the singular rewards—and justifications—for making psychological transformations of data.

The proportional relation between psychological error and psychological magnitudes leads to an important prediction: Thurstone scales should be logarithmically related to magnitude scales. And they are (Galanter and Messick, 1961; Stevens, 1966). Not only is this a gratifying reduction in the number of different psychological functions; it also provides a useful check on scaling techniques. Much of the data collected by Ekman and his colleagues had to do with variables such as seriousness of criminal offenses, prestige of occupations, and political dissatisfaction. The estimates were interesting and of some social utility, and the demonstration that Thurstone scales on these data were always logarithmi-

cally related to the magnitude scales generated confidence in the contextual validity of the scales.

All that is now needed is to show a similar relation between category scales and magnitude scales, and the psychophysicist's house will be in order. The feat would be of especial interest to behaviorists, because the bisection technique is a type of category scale. Nonverbal animals can be trained to bisect continua. (Boakes, 1969; Catania, 1970, p. 9; Herrnstein and van Sommers, 1962), and if the point of bisection could be related to the magnitude (or Thurstone) scale, it would then be a simple task to generate psychological scales for important continua such as time and response rate.

At first inspection, it appeared that category scales might also be logarithmic cousins of magnitude scales. Fechner noted that the photometric intensity of stars was related to their stellar magnitude (a category scale) by a logarithmic function. But Stevens points out an interesting bias resident in that relation. People tend to use the various categories of a scale equally; this "demand characteristic" maximizes the amount of information conveyed to the experimenter, at the cost of biasing their category boundaries away from where they would otherwise be placed. Since there are many more dim stars than bright ones, the stellar magnitude scale is highly biased; laboratory experiments in which this bias is eliminated generate a less curved category scale, one with a logarithmic relation neither to the physical stimulus scale nor to the magnitude scale. To date, no simple relation that holds across studies has been found between category scales and magnitude scales.

The method Stevens developed to eliminate the bias in category scaling is interesting, and worth commenting on because of its general utility. The process is one of iteration: start with an arbitrary distribution of stimuli on the continuum to be rated, and collect ratings. With these, construct a first approximation to a psychological scale. Move the stimuli around so they are equidistant on the "protoscale", and have the stimuli rated again. This time the stimuli will occur approximately equally often in each of the categories, so the bias will be greatly diminished. If necessary, a third iteration will eliminate the bias completely.

Another place where iteration is of use is in the averaging of data. If error is normally distributed around the psychological value of a stimulus, it may be more appropriate to average the psychological values than the physical values. But how do we do that, if we do not yet know the psychophysical function? By approximating it with the data of the experiment (*e.g.*, one using the "method of adjustment"), converting the independent variables to first-approximation psychological variables, averaging these transformed values for a second-approximation scale, and so on. I used a variant of this technique when analyzing choice behavior in concurrent-chain schedules, where it was necessary to determine what fixed-interval schedule would be as reinforcing as a particular variable-interval schedule (Killeen, 1968). I collected preference ratings for various fixed-interval schedules, and did a least-squares interpolation for the point of indifference. These data suggested that the proper psychophysical transformation was reciprocation. I then did another least-squares interpolation, this time between preference and the reciprocals of the fixed-interval values (the immediacy of reinforcement). The agreement between data and theory was improved by this maneuver. Had more precision been necessary, I might have employed a third iteration after transforming the independent variable to the relative immediacy of reinforcement, because that was the function finally advocated.

Are such transformations valid? They make sense, and the threat of censure for *post-hoc* analysis should not scare us away from our responsibility for cleaning up our numbers, for climbing the emerging structure of theory for new perspectives on its empirical foundations. Indeed, this is exactly what is intended when we endorse "functional definitions" of operants, punishers, and reinforcers, with theory—Response Induction and the Law of Effect—providing a viewpoint from which we may discern order in our data. That some recent phenomena (adjunctive behavior, sign-tracking, constraints on learning) fall beyond our theoretical ken speaks less for our inductive/empiric bias than it does against our lack of theoretical development.

Incidental observations and asides leaven *Psychophysics*. How much money would it take to make you twice as happy as would a ten-dollar bill? (About $40, because utility

grows as the square root of amount.) If television screens are gray, how can the bad guys wear black hats? (Sensory inhibition.) How large should dots on a map be to represent doubled population? (The diameter should be 1.6 times as large, because the exponent for area is 0.7.) When should the geometric mean be used? (Most of the time, because variability is often relative, seldom absolute.) What happens to psychophysical functions in the presence of a masker or a contrast stimulus? (They become steeper, reflecting a greater sensitivity to changes in the independent variable. *Cf.* behavioral contrast and concurrent-schedule effects.) Are magnitude estimates affected by the immediate history of estimates? (Yes; despite Stevens' historical indifference to Helson's Adaptation Level Theory, he cites Cross' [1973] demonstration of impressive and orderly context effects.) Is the basic psychometric function used in the determination of thresholds a normal ogive? (No, it is a step-function, or ramp, whose form is distorted into an ogive by noise and by averaging.)

Stevens' work was not uncontested, but he always gave his opponents a good fight. The action, however, is moving away from Stevens' psychophysics. Cliff writes: "A scale cannot stand alone, it must be supported by a network of relations, and the broader and tighter the network, the more confidence there is in the scale" (1973, p. 48). Stevens would of course agree, but this philosophy has led many investigators away from the piecemeal generation of undimensional scales, followed by a test—within or across modalities—of their relation to other scales, and toward a frontal assault on the network. The new psychophysics may be called "functional measurement"; it is similar to our "functional definitions", in that you choose as a scale value for a stimulus exactly that value needed to make two or more scales interrelate in a consistent fashion (just as we choose the label "reinforcer" or "punisher" for a stimulus so as to make the ensuing behavior change consistent with the Law of Effect). Functional measurement was spearheaded by Shepard's (1962) seminal paper on nonmetric multidimensional scaling, and extended by many beautiful elaborations of the basic idea, as well as by independent approaches to functional measurement (*e.g.*, Anderson, 1970). These techniques distill a few items of importance—psychological scales on each of the dimensions and rules of combination or distance functions—from many items of less importance and reliability—the raw data comprising the confusions or ratings of each of the compound stimuli. Schneider (1972) provides an introduction to the paradigm along with a nice demonstration of multidimensional scaling of visual stimuli for the pigeon. It is interesting that most functional measurement techniques introduce a new "brass instrument" to the laboratory—a number-crunching computer that can iterate successive approximations to the optimal solution.

But if Stevens' tactics of research are being bypassed, his philosophy will always be mainline science. He distrusted both theory and data that stood alone, and celebrated transformations of either that would bring them into accord and would leave us thereby more fluent in the language of nature.

REFERENCES

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

Boakes, R. A. The bisection of a brightness interval by pigeons. *Journal of the Experimental Analysis of Behavior,* 1969, **12,** 201-209.

Catania, A. C. Reinforcement schedules and psychophysical judgments. In W. H. Schoenfield (Ed.), *The theory of reinforcement schedule.* New York: Appleton-Century-Crofts, 1970. Pp. 1-42.

Cliff, N. Scaling. *Annual review of psychology,* 1973, **24,** 473-506.

Cross, D. V. Sequential dependencies and regression in psychophysical studies. *Perception and Psychophysics,* 1973, **14,** 547-552.

Day, W. F. Radical behaviorism in reconciliation with phenomenology. *Journal of the Experimental Analysis of Behavior,* 1969, **12,** 315-328.

Ekman, G. Discriminal sensitivity on the subjective continuum. *Acta Psychologica,* 1956, **12,** 233-243.

Ekman, G. Weber's law and related functions. *Journal of Psychology,* 1959, **47,** 343-352.

Forscher, B. K. Chaos in the brickyard. *Science,* 1963, **142,** 339.

Galanter, E. and Messick, S. The relation between category and magnitude scales of loudness. *Psychological Review,* 1961, **38,** 363-372.

Herrnstein, R. J. and Van Sommers, P. Method for sensory scaling with animals. *Science,* 1962, **135,** 40-41.

Jenkins, H. M. Sequential organization in schedules of reinforcement. In W. N. Schoenfeld (Ed.), *The theory of reinforcement schedules.* New York: Appleton-Century-Crofts, 1970. Pp. 63-109.

Killeen, P. On the measurement of reinforcement frequency in the study of preference. *Journal of the Experimental Analysis of Behavior,* 1968, **11,** 263-269.

Krantz, D. H. A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology,* 1972, **9,** 168-199.

MacCorquodale, K. and Meehl, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychological Review,* 1948, **55,** 95-107.

Miller, G. A. Stanley Smith Stevens: 1906-1973. *American Journal of Psychology,* 1974, **87,** 279-288.

Platt, J. R. Strong inference. *Science,* 1964, **146,** 347-353.

Schneider, B. A. Multidimensional scaling of color difference in the pigeon. *Perception and Psychophysics,* 1972, **12,** 373-378.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, I. *Psychometrika,* 1962, **27,** 125-140.

Shepard, R. N. *What does the psychophysicist measure?* Unpublished manuscript, 1966.

Skinner, B. F. *The behavior of organisms.* New York: Appleton-Century-Crofts, 1938.

Skinner, B. F. The flight from the laboratory. In J. T. Wilson *et al.* (Eds.), *Current trends in psychological theory.* Pittsburgh: University of Pittsburgh Press, 1958. Reprinted in A. C. Catania's *Contemporary research in operant behavior.* Glenview: Scott Foresman and Company, 1968.

Stevens, S. S. A metric for the social consensus. *Science,* 1966, **151,** 530-541.

Stevens, S. S. Measurement, statistics, and the schemapiric view. *Science,* 1968, **161,** 849-856.

Teghtsoonian, R. On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review,* 1971, **78,** 71-80.

Zuriff, G. E. A behavioral interpretation of psychophysical scaling. *Behaviorism,* 1972, **1,** 118-133.