12. Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.
13. Stevens, S. S. A scale for the measurement of a psychological magnitude: loudness. *Psychol. Rev.*, 1936, 43, 405–416.
14. Stevens, S. S. On the problem of scales for the measurement of psychological magnitudes. *J. Unif. Sci.*, 1939, 9, 94–99.
15. Stevens, S. S. On the theory of scales of measurement. *Science*, 1946, 103, 677–680.
16. Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (ed.), *Handbook of experimental psychology.* New York: Wiley, 1951, pp. 1–49.
17. Stevens, S. S. The measurement of loudness. *J. Acoust. Soc. Amer.*, 1955, 27, 815–829.
18. Stevens, S. S. On the averaging of data. *Science*, 1955, 121, 113–116.
19. Stevens, S. S. On the psychophysical law. *Psychol. Rev.*, 1957, 64, 153–181.
20. Stevens, S. S., and Galanter, E. H. Ratio scales and category scales for a dozen perceptual continua. *J. Exp. Psychol.*, 1957, 54, 377–411.

S. S. STEVENS

8 Measurement, Statistics and the Schemapiric View

A curious antagonism has sometimes infected the relations between measurement and statistics. What ought to proceed as a pact of mutual assistance has seemed to some authors to justify a feud that centers on the degree of independence of the two domains. Thus Humphreys (1) dispenses praise to a textbook because its authors "do not follow the Stevens dictum concerning the precise relationships between scales of measurement and permissible statistical operations." Since that dictum, so-called, lurks as the *bête noire* behind many recurrent complaints, there is need to reexamine its burden and to ask how measurement and statistics shape up in the scientific process—the schemapiric endeavor in which we invent schematic models to map empirical domains.

In those disciplines where measurement is noisy, uncertain, and difficult, it is only natural that statistics should flourish. Of course, if there were no measurement at all, there would be no statistics. At the other extreme, if accurate measurement were achieved in every inquiry, many of the needs for statistics would vanish. Somewhere between the two extremes of no measurement and perfect measurement, perhaps near the psychosocial-behavioral center of gravity, the ratio of statisticizing to measuring reaches its maximum And that is where we find an acute sensitivity to the suggestion that the type of measurement achieved in an experiment may set bounds on the kinds of statistics that will prove appropriate.

After reviewing the issues Anderson (2) concluded that "the statistical test can hardly be cognizant of the empirical meaning of the numbers with which it deals. Consequently," he continued, "the validity of the statistical inference cannot depend on the type of measuring scale used." This sequitur, if we may call it that, demands scrutiny, for it compresses large issues into a few phrases. Here let me observe merely that, however much we may agree that the statistical test

From *Science*, Vol. 161, 30 August 1968, pp. 849–56.

cannot be cognizant of the empirical meaning of the numbers, the same privilege of ignorance can scarcely be extended to experimenters.

Speaking as a statistician, Savage (3) said, "I know of no reason to limit statistical procedures to those involving arithmetic operations consistent with the scale properties of the observed quantities." A statistician, like a computer, may perhaps feign indifference to the origin of the numbers that enter into a statistical computation, but that indifference is not likely to be shared by the scientist. The man in the laboratory may rather suspect that, if something empirically useful is to emerge in the printout, something empirically meaningful must be programed for the input.

Baker, Hardyck, and Petrinovich (4) summed up the distress: "If Stevens' position is correct, it should be emphasized more intensively; if it is incorrect, something should be done to alleviate the lingering feelings of guilt that plague research workers who deliberately use statistics such as *t* on weak measurements." If it is true that guilt must come before repentence, perhaps the age of statistical indifference to the demands of measurement may be drawing to a close. Whatever the outcome, the foregoing samples of opinion suggest that the relation between statistics and measurement is not a settled issue. Nor is it a simple issue, for it exhibits both theoretical and practical aspects. Moreover, peace is not likely to be restored until both the principles and the pragmatics have been resolved.

The Schemapiric Principle

Although measurement began in the empirical mode, with the accent on the counting of moons and paces and warriors, it was destined in modern times to find itself debated in the formal, schematic, syntactical mode, where models can be made to bristle with symbols. Mathematics, which like logic constitutes a formal endeavor, was not always regarded as an arbitrary construction devoid of substantive content, an adventure of postulate and theorem. In early ages mathematics and empirical measurement were as warp and woof, interpenetrating each other so closely that our ancestors thought it proper to prove arithmetic theorems by resort to counting or to some other act of measurement. The divorce took place only in recent times. And mathematics now enjoys full freedom to "play upon symbols," as Gauss phrased it, with no constraints imposed by the demands of empirical measurement.

So also with other formal or schematic systems. The propositions of a formal logic express tautologies that say nothing about the world of tangible stuff. They are analytic statements, so-called, and they stand apart from the synthetic statements that express facts and relations among empirical objects. There is a useful distinction to be made between the analytic, formal, syntactical propositions of logic and the synthetic, empirical statements of substantive discourse.

Sometimes the line may be hard to draw. Quine (5) the logician denies, in fact, that any sharp demarcation can be certified, and debate on the issue between him and Carnap has reached classic if unresolved proportions. For the scientist, meanwhile, the usefulness of the formal-empirical distinction need not be imperiled by the difficulty of making rigorous decisions in borderline cases. It is useful to distinguish between day and night despite the penumbral passage through twilight. So also is it useful to tune ourselves to distinguish between the formally schematic and the empirical substantive.

Probability exhibits the same double aspect, the same schemapiric nature. Mathematical theories of probability inhabit the formal realm as analytic, tautologous, schematic systems, and they say nothing at all about dice, roulette, or lotteries. On the empirical level, however, we count and tabulate events at the gaming table or in the laboratory and note their relative frequencies. Sometimes the relative frequencies stand in isomorphic relation to some property of a mathematical

model of probability; at other times the observed frequencies exhibit scant accord with "expectations."

Those features of statistics that involve a probabilistic schema provide a further instance of a formal-empirical dichotomy: the distinction between the probability model and the statistical data. E. B. Wilson (6), mathematician and statistician, made the point "that one must distinguish critically between probability as a purely mathematical subject of one sort or another, and statistics which cannot be so regarded." Statistics, of course, is a young discipline—one whose voice changes depending on who speaks for it. Many spokesmen would want to broaden the meaning of statistics to include a formal, mathematical segment.

In another context N. R. Hanson (7) pressed a similar distinction when he said, "Mathematics and physics on this account seem *logically* different disciplines, such that the former can only occasionally solve the latter's problems." Indeed, as Hanson later exclaimed, "Physicists have in unison pronounced, 'Let no man join what nature hath sundered, namely, the *formal creation* of spaces and the physical *description* of bodies.' " Yet it is precisely by way of the proper and judicious joining of the schematic with the empirical that we achieve our beneficial and effective mappings of the universe—the schemapiric mappings known as science. The chronic danger lies in our failure to note the distinction between the map and the terrain, between the simulation and the simulated. The map is an analogue, a schema, a model, a theory. Each of those words has a separate flavor, but they all share a common core of meaning. "Contrary to general belief," wrote Simon and Newell (8), "there is no fundamental, 'in principle,' difference between theories and analogies. All theories are analogies, and all analogies are theories." Indeed, the same can be said for all the other terms that designate the associative binding of schematics to empirics—what I have called the schemapiric bond.

Scales and Invariance

Although it could be otherwise if our choice dictated, most measurement involves the assignment of numbers to aspects of objects or events according to one or another rule or convention. The variety of rules invented thus far for the assignment of numbers has already grown enormous, and novel means of measuring continue to emerge. It has proved possible, however, to formulate an invariance criterion for the classification of scales of measurement (9). The resulting systemization of scale types has found uses in contexts ranging from physics (10) to the social sciences (11), but the conception has not enjoyed immunity from criticism (12).

Let me sketch the theory. It can be done very briefly, because details are given in other places (13). The theory proposes that a scale type is defined by the group of transformations under which the scale form remains invariant, as follows.

A *nominal scale* admits any one-to-one substitution of the assigned numbers. Example of a nominal scale: the numbering of football players.

An *ordinal scale* can be transformed by any increasing monotonic function. Example of an ordinal scale: the hardness scale determined by the ability of one mineral to scratch another.

An *interval scale* can be subjected to a linear transformation. Examples of interval scales: temperature Fahrenheit and Celsius, calendar time, potential energy.

A *ratio scale* admits only multiplication by a constant. Examples of ratio scales: length, weight, density, temperature Kelvin, time intervals, loudness in sones.

The foregoing scales represent the four types in common use. Other types are possible. The permissible transformations defining a scale type are those that keep intact the empirical information depicted by the scale. If the empirical information has been preserved, the scale form is said to remain invariant. The critical isomorphism is main-

Table 1. Examples of statistical measures appropriate to measurements made on various types of scales. The scale type is defined by the manner in which scale numbers can be transformed without the loss of empirical information. The statistical measures listed are those that remain invariant, as regards either value or reference, under the transformations allowed by the scale type.

Scale type	*Measures of location*	*Dispersion*	*Association or correlation*	*Significance tests*
Nominal	Mode	Information *H*	Information transmitted *T*	Chi square Fisher's exact test
Ordinal	Median	Percentiles	Rank correlation	Sign test Run test
Interval	Arithmetic mean	Standard deviation Average deviation	Product-moment correlation Correlation ratio	*t* test *F* test
Ratio	Geometric mean Harmonic mean	Percent variation Decilog dispersion		

tained. That indeed is the principle of invariance that lies at the heart of the conception. More formal presentations of the foregoing theory have been undertaken by other authors, a recent one, for example, by Lea (14).

Unfortunately, those who demand an abstract tidiness that is completely aseptic may demur at the thought that the decision whether a particular scale enjoys the privilege of a particular transformation group depends on something so ill defined as the preservation of empirical information. For one thing, an empirical operation is always attended by error. Thus Lebesgue (15), who strove so well to perfect the concept of mathematical measure, took explicit note that, in the assignment of number to a physical magnitude, precision can be pushed, as he said, "in actuality only up to a certain error. It never enables us," he continued, "to discriminate between one number and all the numbers that are extremely close to it."

A second disconcerting feature of the invariance criterion lies in the difficulty of specifying the empirical information that is to be preserved. What can it be other than the information that we think we have captured by creating the scale in the first place? We may, for example, perform operations that allow us simply to identify or discriminate a particular property of an object. Sometimes we want to preserve nothing more than that simple outcome, the identification or nominal classification of the items of interest. Or we may go further, provided our empirical operations permit, and determine rank orders, equal intervals, or equal ratios. If we want our number assignments to reflect one or another accrual in information, we are free to transform the scale numbers only in a way that does not lose or distort the desired information. The choice remains ours.

Although some writers have found it possible to read an element of prescription—even proscription—into the invariance principle, as a systematizing device the principle contains no normative force. It can be read more as a description of the obvious than as a directive. It says that, once an isomorphism has been mapped out between aspects of objects or events, on the one hand, and some one or more features of the number system, on the other hand, the isomorphism can be upset by whatever transformations fail to preserve it. Precisely what is preserved or not preserved in a particular circumstance depends upon the empirical operations. Since actual day-to-day measurements range from

muddled to meticulous, our ability to classify them in terms of scale type must range from hopelessly uncertain to relatively secure.

The group invariance that defines a scale type serves in turn to delimit the statistical procedures that can be said to be appropriate to a given measurement scale (16). Examples of appropriate statistics are tabulated in Table 1. Under the permissible transformations of a measurement scale, some appropriate statistics remain invariant in value (example: the correlation coefficient r keeps its value under linear transformations). Other statistics change value but refer to the same item or location (example: the median changes its value but continues to refer to mid-distribution under ordinal transformations).

Reconciliation and New Problems

Two developments may serve to ease the apprehension among those who may have felt threatened by a theory of measurement that seems to place bounds on our freedom to calculate. One is a clearer understanding of the bipartite, schemapiric nature of the scientific enterprise. When the issue concerns only the schema—when, for example, critical ratios are calculated for an assumed binomial distribution—then indeed it is purely a matter of relations within a mathematical model. Natural facts stand silent. Empirical considerations impose no constraints. When, however, the text asserts a relation among such things as measured differences or variabilities, we have a right and an obligation to inquire about the operations that underlie the measurements. Those operations determine, in turn, the type of scale achieved.

The two-part schemapiric view was expressed by Hays (17) in a much-praised book: "If the statistical method involves the procedures of arithmetic used on numerical scores, then the numerical answer is formally correct. . . . The difficulty comes with the interpretation of these numbers back into statements about the real world. If nonsense is put into the mathematical system, nonsense is sure to come out."

At the level of the formal model, then, statistical computations may proceed as freely as in any other syntactical exercise, unimpeded by any material outcome of empirical measurement. Nor does measurement have a presumptive voice in the creation of the statistical models themselves. As Hogben (18) said in his forthright dissection of statistical theory, "It is entirely defensible to formulate an axiomatic approach to the theory of probability as an internally consistent set of propositions, if one is content to leave to those in closer contact with reality the last word on the usefulness of the outcome." Both Hays and Hogben insist that the user of statistics, the man in the laboratory, the maker of measurements, must decide the meaning of the numbers and their capacity to advance empirical inquiry.

The second road to reconciliation winds through a region only partly explored, a region wherein lies the pragmatic problem of appraising the wages of transgression. What is the degree of risk entailed when use is made of statistics that may be inappropriate in the strict sense that they fail the test of invariance under permissible scale transformations? Specifically, let us assume that a set of items can be set in rank order, but, by the operations thus far invented, distances between the items cannot be determined. We have an ordinal but not an interval scale. What happens then if interval-scale statistics are applied to the ordinally scaled items? Therein lies a question of first-rate substance and one that should be amenable to unemotional investigation. It promises well that a few answers have already been forthcoming.

First there is the oft-heeded counsel of common sense. In the averaging of test scores, says Mosteller (19), "It seems sensible to use the statistics appropriate to the type of scale I think I am near. In taking such action we may find the justification vague and fuzzy. One reason for this vagueness is that we have not yet studied enough about classes of scales,

classes appropriate to real life measurement, with perhaps real life bias and error variance."

How some of the vagueness of which Mosteller spoke can perhaps be removed is illustrated by the study of Abelson and Tukey (20) who showed how bounds may be determined for the risk involved when an interval-scale statistic is used with an ordinal scale. Specifically, they explored the effect on r^2 of a game against nature in which nature does its best (or worst!) to minimize the value of r^2. In this game of regression analysis, many interesting cases were explored, but, as the authors said, their methods need extension to other cases. They noted that we often know more about ordinal data than mere rank order. We may have reason to believe, they said, "that the scale is no worse than mildly curvilinear, that Nature behaves smoothly in some sense." Indeed the continued use of parametric statistics with ordinal data rests on that belief, a belief sustained in large measure by the pragmatic usefulness of the results achieved.

In a more synthetic study than the foregoing analysis, Baker *et al.* (4) imposed sets of monotonic transformations on an assumed set of data, and calculated the effect on the t distribution. The purpose was to compare distributions of t for data drawn from an equal-interval scale with distributions of t for several types of assumed distortions of the equal intervals. By and large, the effects on the computed t distributions were not large, and the authors concluded "that strong statistics such as the t test are more than adequate to cope with weak [ordinal] measurements...". It should be noted, however, that the values of t were affected by the nonlinear transformations. As the authors said, "The correspondence between values of t based on the criterion unit interval scores and values of t based on [nonlinear] transformations decreases regularly and dramatically...as the departure from linear transformations becomes more extreme."

Whatever the substantive outcome of such investigations may prove to be, they point the way to reconciliation through orderly inquiry. Debate gives way to calculation. The question is thereby made to turn, not on whether the measurement scale determines the choice of a statistical procedure, but on how and to what degree an inappropriate statistic may lead to a deviant conclusion. The solution of such problems may help to refurbish the complexion of measurement theory, which has been accused of proscribing those statistics that do not remain invariant under the transformations appropriate to a given scale. By spelling out the costs, we may convert the issue from a seeming proscription to a calculated risk.

The type of measurement achieved is not, of course, the only consideration affecting the applicability of parametric statistics. Bradley is one of many scholars who have sifted the consequences of violating the assumptions that underlie some of the common parametric tests (21). As one outcome of his studies, Bradley concluded, "The contention that, when its assumptions are violated, a parametric test is still to be preferred to a distribution-free test because it is 'more efficient' is therefore a monumental *non sequitur*. The point is not at all academic...violations in a test's assumptions may be attended by profound changes in its power." That conclusion is not without relevance to scales of measurement, for when ordinal data are forced into the equal-interval mold, parametric assumptions are apt to be violated. It is then that a so-called distribution-free statistic may prove more efficient than its parametric counterpart.

Although better accommodation among certain of the contending statistical usages may be brought about by computer-aided studies, there remain many statistics that find their use only with specific kinds of scales. A single example may suffice. In a classic text-book, written with a captivating clarity, Peters and Van Voorhis (22) got hung up on a minor point concerning the procedure to be used in comparing variabilities. They noted that Karl Pearson had proposed a measure called the coefficient of

variation, which expresses the standard deviation as a percentage of the mean. The authors expressed doubts about its value, however, because it tells "more about the extent to which the scores are padded by a dislocation of the zero point than it does about comparable variabilities." The examples and arguments given by the authors make it plain that the coefficient of variation has little business being used with what I have called interval scales. But since their book antedated my publication in 1946 of the defining invariances for interval and ratio scales, Peters and Van Voorhis did not have a convenient way to state the relationship made explicit in Table 1, namely, that the coefficient of variation, being itself a ratio, called for a ratio scale.

Complexities and Pitfalls

Concepts like relative variability have the virtue of being uncomplicated and easy for the scientist to grasp. They fit his idiom. But in the current statistics explosion, which showers the investigator with a dense fallout of new statistical models, the scientist is likely to lose the thread on many issues. It is then that the theory of measurement, with an anchor hooked fast in empirical reality, may serve as a sanctuary against the turbulence of specialized abstraction.

"As a mathematical discipline travels far from its empirical source," said von Neumann (23), "there is grave danger that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganized mass of details and complexities." He went on to say that, "After much 'abstract' inbreeding, a mathematical subject is in danger of degeneration. At the inception the style is usually classical; when it shows signs of becoming baroque, then the danger signal is up."

There is a sense, one suspects, in which statistics needs measurement more than measurement needs statistics. R. A. Fisher alluded to that need in his discourse on the nature of probability (24). "I am quite sure," he said, "it is only personal contact with the business of the improvement of natural knowledge in the natural sciences that is capable to keep straight the thought of mathematically-minded people who have to grope their way through the complex entanglements of error...."

And lest the physical sciences should seem immune to what Schwartz (25) called "the pernicious influence of mathematics," consider his diagnosis: "Thus, in its relations with science, mathematics depends on an intellectual effort outside of mathematics for the crucial specification of the approximation which mathematics is to take literally. Give a mathematician a situation which is the least bit ill-defined—he will first of all make it well defined. Perhaps appropriately, but perhaps also inappropriately....That form of wisdom which is the opposite of single-mindedness, the ability to keep many threads in hand, to draw for an argument from many disparate sources, is quite foreign to mathematics.... Quite typically, science leaps ahead and mathematics plods behind."

Progress in statistics often follows a similar road from practice to prescription—from field trials to the formalization of principles. As Kruskal (26) said "Theoretical study of a statistical procedure often comes after its intuitive proposal and use." Unfortunately for the empirical concerns of the practitioners, however, there is, as Kruskal added, "almost no end to the possible theoretical study of even the simplest procedure." So the discipline wanders far from its empirical source, and form loses sight of substance.

Not only do the forward thrusts of science often precede the mopping-up campaigns of the mathematical schema builders, but measurement itself may often find implementation only after some basic conception has been voiced. Textbooks, those distilled artifices of science, like to picture scientific conceptions as built on measurement, but the working scientist is more apt to devise his measure-

ments to suit his conceptions. As Kuhn (27) said, "The route from theory or law to measurement can almost never be travelled backwards. Numbers gathered without some knowledge of the regularity to be expected almost never to speak for themselves. Almost certainly they remain just numbers." Yet who would deny that some ears, more tuned to numbers, may hear them speak in fresh and revealing ways?

The intent here is not, of course, to affront the qualities of a discipline as useful as mathematics. Its virtues and power are too great to need extolling, but in power lies a certain danger. For mathematics, like a computer, obeys commands and asks no questions. It will process any input, however devoid of scientific sense, and it will bedeck in formulas both the meaningful and the absurd. In the behavioral sciences, where the discernment for nonsense is perhaps less sharply honed than in the physical sciences, the vigil must remain especially alert against the intrusion of a defective theory merely because it carries a mathematical visa. An absurdity in full formularized attire may be more seductive than an absurdity undressed.

Distributions and Decisions

The scientist often scales items, counts them, and plots their frequency distributions. He is sometimes interested in the form of such distributions. If his data have been obtained from measurements made on interval or ratio scales, the shape of the distribution stays put (up to a scale factor) under those transformations that are permissible, namely, those that preserve the empirical information contained in the measurements. The principle seems straightforward. But what happens when the state of the art can produce no more than a rank ordering, and hence nothing better than an ordinal scale? The abscissa of the frequency distribution then loses its metric meaning and becomes like a rubber band, capable of all sorts of monotonic stretchings. With each non-linear transformation of the scale, the form of the distribution changes. Thereupon the distribution loses structure, and we find it futile to ask whether the shape approximates a particular form, whether normal, rectangular, or whatever.

Working on the formal level, the statistician may contrive a schematic model by first assuming a frequency function, or a distribution function, of one kind or another. At the abstract level of mathematical creation, there can, of course, be no quarrel with the statistician's approach to his task. The caution light turns on, however, as soon as the model is asked to mirror an empirical domain. We must then invoke a set of semantic rules—coordinating definitions—in order to identify correspondences between model and reality. What shall we say about the frequency function $f(x)$ when the problem before us allows only an ordinal scale? Shall x be subject to a nonlinear transformation after $f(x)$ has been specified? If so, what does the transformation do to the model and to the predictions it forecasts?

The scientist has reason to feel that a statistical model that specifies the form of a canonical distribution becomes uninterpretable when the empirical domain concerns only ordinal data. Yet many consumers of statistics seem to disregard what to others is a rather obvious and critical problem. Thus Burke (28) proposed to draw "two random samples from populations known to be normal" and then "to test the hypothesis that the two populations have the same mean . . . under the assumption that the scale is ordinal at best." How, we must ask, can normality be known when only order can be certified?

The assumption of normality is repeated so blithely and so often that it becomes a kind of incantation. If enough of us sin, perhaps transgression becomes a virtue. But in the instance before us, where the numbers to be fed into the statistical mill result from operations that allow only a rank ordering, maybe we have gone too far. Consider a permissible transformation. Let us cube all the numbers. The rank order would stand as before. But

what do we then say about normality? If we can know nothing about the intervals on the scale of a variable, the postulation that a distribution has a particular form would appear to proclaim a hope, not a circumstance.

The assertion that a variable is normally distributed when the variable is amenable only to ordinal measurement may loom as an acute contradiction, but it qualifies as neither the worst nor the most frequent infraction by some of the practitioners of hypothesis testing. Scientific decision by statistical calculation has become the common mode in many behavioral disciplines. In six psychological journals (29), for example, the proportion of articles that employed one or another kind of inferential statistic rose steadily from 56 percent in 1948 to 91 percent in 1962. In the *Journal of Educational Psychology* the proportion rose from 36 to 100 percent.

What does it mean? Can no one recognize a decisive result without a significance test? How much can the burgeoning of computation be blamed on fad? How often does inferential computation serve as a premature excuse for going to press? Whether the scholar has discovered something or not, he can sometimes subject his data to an analysis of variance, a *t* test, or some other device that will produce a so-called objective measure of "significance." The illusion of objectivity seems to preserve itself despite the admitted necessity for the investigator to make improbable assumptions, and to pluck off the top of his head a figure for the level of probability that he will consider significant. His argument that convention has already chosen the level that he will use does not quite absolve him.

Lubin (30) has a name for those who censure the computational and applaud the experimental in the search for scientific certainty. He calls them stochastophobes. An apt title, if applied to those whose eagerness to lay hold on the natural fact may generate impatience at the gratuitous processing of data. The extreme stochastophobe is likely to ask: What scientific discoveries owe their existence to the techniques of statistical analysis or inference? If exercises in statistical inference have occasioned few instances of a scientific breakthrough, the stochastophobe may want to ask by what magical view the stochastophile perceives glamour in statistics. The charm may stem in part from the prestige that mathematics, however inapposite, confers on those who display the dexterity of calculation. For some stochastophiles the appeal may have no deeper roots than a preference for the prudent posture at a desk as opposed to the harsher, more venturesome stance in the field or the laboratory.

The aspersions voiced by stochastophobes fall mainly on those scientists who seem, by the surfeit of their statistical chants, to turn data treatment into hierurgy. These are not the statisticians themselves, for they see statistics for what it is, a straightforward discipline designed to amplify the power of common sense in the discernment of order amid complexity. By showing how to amend the mismatch in the impedance between question and evidence, the statistician improves the probability that our experiments will speak of greater clarity. And by weighing the entailments of relevant assumptions, he shows us how to milk the most from some of those fortuitous experiments that nature performs once and may never perform again. The stochastophobe should find no quarrel here. Rather he should turn his despair into a hope that the problem of the relevance of this or that statistical model may lead the research man toward thoughtful inquiry, not to a reflex decision based on a burst of computation.

Measurement

If the vehemence of the debate that centers on the nature and conditions of statistical inference has hinted at the vulnerability of the conception, what can be said about the other partner in the enterprise? Is the theory of measurement a settled matter? Apparently not, for it remains a topic of trenchant

inquiry, not yet ready to rest its case. And debate continues.

The typical scientist pays little attention to the theory of measurement, and with good reason, for the laboratory procedures for most measurements have been well worked out, and the scientist knows how to read his dials. Most of the variables are measured on well-defined, well-instrumented ratio scales.

Among those whose interests center on variables that are not reducible to meter readings, however, the concern with measurement stays acute. How, for example, shall we measure subjective value (what the economists calls utility), or perceived brightness, or the seriousness of crimes? Those are some of the substantive problems that have forced a revision in our approach to measurement. They have entailed a loosening of the restricted view bequeathed us by the tradition of Helmholtz and Campbell—the view that the axioms of additivity must govern what we call measurement (31). As a related development, new axiomatic systems have appeared, including axioms by Luce and Tukey (32) for a novel "conjoint" approach to fundamental measurement. But the purpose here is not to survey the formal, schematic models that have flowered in the various sciences, for the practice and conception of measurement has as yet been little influenced by them.

As with many syntactical developments, measurement models sometimes drift off into the vacuum of abstraction and become decoupled from their concrete reference. Even those authors who freely admit the empirical features as partners in the formulation of measurement may find themselves seeming to downgrade the empirical in favor of the formal. Thus we find Suppes and Zinnes (33) saying, "Some writers . . . appear to define scales in terms of the existence of certain empirical operations. . . . In the present formulation of scale type, no mention is made of the kinds of 'direct' observations or empirical relations that exist. . . . Precisely what empirical operations are involved in the empricial system is of no consequence."

How then do we distinguish different types of scales? How, in particular, do we know whether a given scale belongs among the interval scales? Suppes and Zinnes gave what I think is a proper answer: "We ask if all the admissible numerical assignments are related by a linear transformation." That, however, is not a complete answer. There remains a further question: What is it that makes a class of numerical assignments admissible? A full theory of measurement cannot detach itself from the empirical substrate that gives it meaning. But the theorist grows impatient with the empirical lumps that ruffle the fine laminar flow within his models just as the laboratory fellow may disdain the arid swirls of hieroglyphics that pose as paradigms of his measurements.

Although a congenial conciliation between those two polar temperaments, the modeler and the measurer, may lie beyond reasonable expectations, a tempering détente may prove viable. The two components of schemapirics must both be accredited, each in its own imperative role. To the understanding of the world about us, neither the formal model nor the concrete measure is dispensable.

Matching and Mapping

Instead of starting with origins, many accounts of measurement begin with one or another advanced state of the measuring process, a state in which units and metrics can be taken for granted. At that level, the topic already has the crust of convention upon it, obscuring the deeper problems related to its nature.

If we try to push the problem of measurement back closer to its primordial operations, we find, I think, that the basic operation is always a process of matching. That statement may sound innocent enough, but it contains a useful prescription. It suggests, for example, that if you would understand the essence of a given measuring procedure, you should ask what was matched to what. If the query leads to a pointer reading, do not stop there; ask the same question about the calibration procedure

that was applied to the instruments anterior to the pointer: What was matched to what? Diligent pursuit of that question along the chain of measuring operations leads to some of the elemental operations of science.

Or we may start nearer the primordium. The sketchiness of the record forces us to conjecture the earliest history, but quite probably our forefather kept score on the numerosity of his possessions with the aid of piles of pebbles (Latin: *calculi*) or by means of some other tallying device. He paired off items against pebbles by means of a primitive matching operation, and he thereby measured his hoard.

Let us pause at this point to consider the preceding clause. Can the ancestor in question be said to have measured his possessions if he had no number system? Not if we insist on taking literally the definition often given, namely, that measurement is the assignment of numbers to objects or events according to rule. This definition serves a good purpose in many contexts, but it presumes a stage of development beyond the one that we are now seeking to probe. In an elemental sense, the matching or assigning of numbers is a sufficient but not a necessary condition for measurement, for other kinds of matching may give measures.

Numbers presumably arose after our ancestor invented names for the collection of pebbles, or perhaps for the more convenient collections, the fingers. He could then match name to collection, and collection to possessions. That gave him a method of counting, for, by pairing off each item against a finger name in an order decided upon, the name of the collection of items, and hence the numerosity of the items, was specified.

The matching principle leads to the concept of cardinality. Two sets have the same cardinal number if they can be paired off in one-to-one relation to each other. By itself, this cardinal pairing off says nothing about order. (Dictionaries often disagree with the mathematicians on the definition of cardinality, but the mathematical usage recommends itself here.) We find the cardinal principle embodied in the symbols used for the numerals in many forms of writing. Thus the Roman numeral VI pictures a hand V and a finger I.

Let us return again to our central question. In the early cardinal procedure of matching item to item, fingers to items, or names to items, at what point shall we say that measurement began? Perhaps we had best not seek a line of demarcation between measurement and matching. It may be better to go all the way and propose an unstinted definition as follows: Measurement is the matching of an aspect of one domain to an aspect of another.

The operation of matching eventuates, of course, in one domain's being mapped into another, as regards one or more attributes of the two domains. In the larger sense, then, whenever a feature of one domain is mapped isomorphically in some relation with a feature of another domain, measurement is achieved. The relation is potentially symmetrical. Our hypothetical forefather could measure his collection of fish by means of his pile of pebbles, or his pile of pebbles by means of his collection of fish.

Our contemporary concern lies not, of course, with pebbles and fish, but with a principle. We need to break the hull that confines the custom of our thought about these matters. The concern is more than merely academic, however, especially in the field of psychophysics. One justification for the enlarged view of measurement lies in a development in sensory measurement known as cross-modality matching (34). In a suitable laboratory setup, the subject is asked, for example, to adjust the loudness of a sound applied to his ears in order to make it seem equal to the perceived strength of a vibration applied to his finger. The amplitude of the vibration is then changed and the matching process is repeated. An equal sensation function is thereby mapped out, as illustrated in Fig. 1. Loudness has been matched in that manner to ranges of values on some ten other perceptual continua, always with the result that the matching function approximates a power function (35). In other words, in order to produce equal apparent

intensity, the amplitude of the sound p must be a power function of the amplitude of the vibration a, or $p = a^b$, where b is the exponent. Or, more simply, the logarithms of the stimuli are linearly related, which means that ratios of stimuli are proportional.

Experiments suggest that the power function obtains between all pairs of intensive perceptual continua, and that the matchings exhibit a strong degree of transitivity in the sense that the exponents form an interconnected net. If two matching functions have one continuum in common, we can predict fairly well the exponent of the matching function between the other two continua.

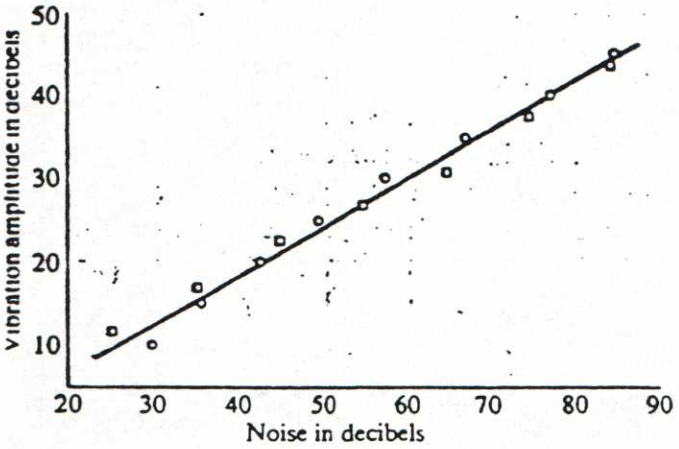

Fig. 1. Equal-sensation function for cross-modality matching between loudness and vibration. The squares indicate that the observers adjusted the intensity of vibration on the fingertip to match the loudness of a noise delivered by earphones. The circles indicate that the observers adjusted the loudness to match the vibration. Each point is the decibel average of 20 matches, two by each often observers. Since the coordinates are logarithmic, the straight line indicates a power function.

Now, once we have mapped out the matching function between loudness and vibration, we can, if we choose, measure the subjective strength of the vibration in terms of its equivalent loudness. Or, more generally, if all pairs of continua have been matched, we can select any one continuum to serve as the reference continuum in terms of which we then measure the subjective magnitude on each of the other continua.

In the description of a measurement system that rests on cross-modality matching, no mention has been made of numbers. If we are willing to start from scratch in a measurement of this kind, numbers can in principle be dispensed with. They would, to be sure, have practical uses in the conduct of the experiments, but by using other signs or tokens to identify the stimuli we could presumably eliminate numbers completely. It would be a tour de force, no doubt, but an instructive one.

Instead of dispensing with numbers, the practice in many psychophysical studies has been to treat numbers as one of the perceptual continua in the cross-modality matching experiment. Thus in what has come to be known as the method of magnitude estimation, numbers are matched to loudness, say. In the reverse procedure, called magnitude production, the subject adjusts the loudness to match a series of numbers given by the experimenter (36). And as might be expected, despite all the other kinds of cross-modality matches that have been made, it is the number continuum that most authors select as the reference continuum (exponent = 1.0) in terms of which the exponent values for the other perceptual continua are stated. But the point deserves to be stressed: the choice of number as the reference continuum is wholly arbitrary, albeit eminently convenient.

Summary

Back in the days when measurement meant mainly counting, and statistics meant mainly the inventory of the state, the simple descriptive procedures of enumeration and averaging occasioned minimum conflict between measurement and statistics. But as measurement pushed on into novel behavioral domains, and statistics turned to the formalizing of stochastic models, the one-time intimate relation between the two activities dissolved into occasional misunderstanding. Measurement and statistics must live in peace, however, for both must participate in the schemapiric

enterprise by which the schematic model is made to map the empirical observation.

Science presents itself as a two-faced, bipartite endeavor looking at once toward the formal, analytic, schematic features of model-building, and toward the concrete, empirical, experiential observations by which we test the usefulness of a particular representation. Schematics and empirics are both essential to science, and full understanding demands that we know which is which.

Measurement provides the numbers that enter the statistical table. But the numbers that issue from measurements have strings attached, for they carry the imprint of the operations by which they were obtained. Some transformations on the numbers will leave intact the information gained by the measurements; other transformations will destroy the desired isomorphism between the measurement scale and the property assessed. Scales of measurement therefore find a useful classification on the basis of a principle of invariance: each of the common scale types (nominal, ordinal, interval, and ratio) is defined by a group of transformations that leaves a particular isomorphism unimpaired.

Since the transformations allowed by a given scale type will alter the numbers that enter into a statistical procedure, the procedure ought properly to be one that can withstand that particular kind of number alteration. Therein lies the primacy of measurement: it sets bounds on the appropriateness of statistical operations. The widespread use on ordinal scales of statistics appropriate only to interval or ratio scales can be said to violate a technical canon, but in many instances the outcome has demonstrable utility. A few workers have begun to assess the degree of risk entailed by the use of statistics that do not remain invariant under the permissible scale transformations.

The view is proposed that measurement can be most liberally construed as the process of matching elements of one domain to those of another domain. In most kinds of measurement we match numbers to objects or events, but other matchings have been found to serve a useful purpose. The cross-modality matching of one sensory continuum to another has shown that sensory intensity increases as the stimulus intensity raised to a power. The generality of that finding supports a psychophysical law expressible as a simple invariance: equal stimulus ratios produce equal sensation ratios.

References

1. L. Humphreys, *Contemp. Psychol.* 9, 76 (1964).
2. N. H. Anderson, *Psychol. Bull.* 58, 305 (1961).
3. I. R. Savage, *J. Amer. Statist. Ass.* 52, 331 (1957).
4. B. O. Baker, C. D. Hardyck, L. F. Petrinovich, *Educ. Psychol. Meas.* 26, 291 (1966).
5. W. V. O. Quine, *The Ways of Paradox and Other Essays* (Random House, New York, 1966), pp. 126–134.
6. E. B. Wilson, *Proc. Natl. Acad. Sci. U.S.* 51, 539 (1964).
7. N. R. Hanson, *Philos. Sci.* 30, 107 (1963).
8. H. A. Simon and A. Newell, in *The State of the Social Sciences*, L. D. White, Ed. (Univ. of Chicago Press, Chicago, 1956), pp. 66–83.
9. S. S. Stevens, *Science* 103, 677 (1946).
10. F. B. Silsbee, *J. Wash. Acad. Sci.* 41, 213 (1951).
11. B. F. Green, in *Handbook of Social Psychology*, G. Lindzey, Ed. (Addison-Wesley, Reading, Mass, 1954), pp. 335–369.
12. Among those who have commented are B. Ellis, *Basic Concepts of Measurement* (University Press, Cambridge, England, 1966); B. Grunstra, "On Distinguishing Types of Measurement," *Boston Studies Phil. Sci.*, vol. 4 (Humanities Press, in press); S. Ross, *Logical Foundations of Psychological Measurement* (Scandinavian University Books, Munksgaard, Copenhagen, 1964); W. W. Rozeboom, *Synthese* 16, 170–233 (1966); W. S. Torgerson, *Theory and Methods of Scaling* (Wiley, New York, 1958).
13. S. S. Stevens, in *Handbook of Experimental Psychology*, S. S. Stevens, Ed. (Wiley, New York, 1951), pp. 1–49; in *Measurement: Definitions and Theories*, C. W. Churchman and P. Ratoosh, Eds. (Wiley, New York, 1959), pp. 18–64.
14. W. A. Lea, "A Formalization of Measurement Scale Forms" (Technical Memo. KC-T-024, Computer Research Lab., NASA Electronics Res. Ctr., Cambridge, Mass., June 1967).

15. H. Lebesgue, *Measure and the Integral*, K. O. May, Ed. (Holden-Day, San Francisco, 1966).
16. Other summarizing tables are presented by V. Senders, *Measurement and Statistics* (Oxford Univ. Press, New York, 1958). A further analysis of appropriate statistics has been presented by E. W. Adams, R. F. Fagot, R. E. Robinson, *Psychometrika* 30, 99 (1965).
17. W. L. Hays, *Statistics for Psychologists* (Holt, Rinehart & Winston, New York, 1963).
18. L. Hogben, *Statistical Theory* (Norton, New York, 1958.)
19. F. Mosteller, *Psychometrika* 23, 279 (1958).
20. R. P. Abelson and J. W. Tukey, *Efficient Conversion of Non-Metric Information into Metric Information* (Amer. Statist. Ass., Social Statist. Sec., December 1959), pp. 226–230; see also *Ann. Math. Stat.* 34, 1347 (1963).
21. J. V. Bradley, "Studies in Research Methodology: II. Consequences of Violating Parametric Assumptions—Facts and Fallacy" (WADC Tech. Rep. 58-574 [II]. Aerospace Med. Lab., Wright-Patterson AFB, Ohio, September 1959).
22. C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases* (McGraw-Hill, New York, 1940).
23. J. von Neumann, in *The Works of the Mind*, R. B. Heywood, Ed. (Univ. of Chicago Press, Chicago, 1947), pp. 180–196.
24. R. A. Fisher, *Smoking, the Cancer Controversy* (Oliver and Boyd, Edinburgh, 1959).
25. J. Schwartz, in *Logic, Methodology and Philosophy of Science*, E. Nagal *et al.*, Eds., (Stanford Univ. Press, Stanford, Calif., 1962), pp. 356–360.
26. W. R. Kruskal, in *International Encyclopedia of the Social Sciences* (Macmillan and Free Press, New York, 1968), vol. 15, pp. 206–224.
27. T. S. Kuln, in *Quantification*, H. Woolf, Ed. (Bobbs-Merrill, Indianapolis, Ind., 1961), pp. 31–63.
28. C. J. Burke, in *Theories in Contemporary Psychology*, M. H. Marx, Ed. (Macmillan, New York, 1963), pp. 147–159.
29. The journals were tabulated by E. S. Edgington, *Amer. Psychologist* 19, 202 (1964); also personal communication.
30. A. Lubin, in *Annual Review of Psychology* (Annual Reviews, Palo Alto, Calif., 1962), vol. 13, pp. 345–370.
31. H. v. Helmholtz, "Zählen und Messen," in *Philosophische Aufsätze* (Fues's Verlag, Leipzig, 1887), pp. 17–52; N. R. Campbell, *Physics: the Elements* [1920] (reissued as *The Philosophy of Theory and Experiment* by Dover, New York, 1957; *Symposium: Measurement and its Importance for Philosophy*, Aristotelian Soc., suppl., vol. 17 (Harrison and Sons, London, 1938).
32. R. D. Luce and J. W. Tukey, *J. Math. Psychol.* 1, 1 (1964).
33. P. Suppes and J. L. Zinnes, in *Handbook of Mathematical Psychology*, R. D. Luce *et al.*, Eds. (Wiley, New York, 1963), pp. 1–76.
34. S. S. Stevens, *J. Exp. Psychol.* 57, 201 (1959); *Amer. Sci.* 54, 385 (1966).
35. S. S. Stevens, *Percept. Psychophys.* 1, 5 (1966).
36. S. S. Stevens and H. B. Greenbaum, *ibid.*, p. 439.
37. This article (Laboratory of Psychophysics Rept. PPR-336-118) was prepared with support from NIH grant NB-02974 and NSF grant GB-3211.