Clyde H. Coombs, THE UNIVERSITY OF MICHIGAN

A THEORY OF DATA

MATHESIS PRESS • 2664 LOWELL RD. • ANN ARBOR, MICH. • 48103

PART 1

BASIC CONCEPTS

Data may be viewed as relations between points in a space. This geometric viewpoint, which we shall develop and explore, leads to a simple classification of types of data. One consequence of this classification is the development of new scaling models; another is the highlighting of certain similarities among data and models from different areas in psychology.

Other viewpoints are possible. For the purposes of this book, however, we use the term data in this restricted geometric sense.

The material contained in this part is generally prerequisite to most of the rest of the book, in that the concepts and vocabulary introduced in these chapters are used freely throughout the remainder. The first chapter is designed to give an overview of this theory of data and explains the organization of the book. The second chapter assembles in one unit a variety of related material on collecting and analyzing data relevant to each of the several parts of the book that follow.

CHAPTER 1

An Overview of a Theory of Data

This book is a report of fourteen years of research into the foundations of psychological measurement. Our orientation to the subject is basically geometrical, and from this viewpoint we find that psychological measurement models, normally dressed in specific and different behavioral languages, may reveal interesting and suggestive interrelations when perceived in a common abstract language.

This consideration of data as relations between points in space leads to a simple classification system of the basic kinds of data which should prove useful (1) to the teacher of psychological measurement and scaling, (2) to the experimenter deciding what method to use in collecting and analyzing data, and (3) to the theoretician in psychological measurement.

To the teacher the value lies in the order and structure introduced to a field which has been developing very rapidly and somewhat chaotically. The similarities among scaling methods are a major feature of this system, and it thereby facilitates learning through generalization and transfer. To the experimenter the value lies in the logical structure placed on methods of collecting and analyzing data, in that the repertoire is increased and some criteria for evaluation and selection are provided. To the theoretician the value lies in the generation of new problems at the same time that possible lines of solution are suggested.

1. SCOPE OF THE THEORY

Behavioral scientists follow a great variety of methods under the general rubrics of collecting and analyzing data. We propose to introduce order into this abundance by formulating in a universal language the

processes by which behavioral data are made, of what they are made, and how they become measurements. In the course of doing this the term *data* itself will take on a restricted (not *different*) meaning, and we become persuaded that the data are in part a product of the mind of the observer.

The restricted meaning that is given here to the term *data* arises from the fact that it has two common uses in behavioral science. The term is commonly used to refer both to the recorded observations and to that which is analyzed. These are not necessarily the same thing, and a distinction is imperative. This distinction seems a subtle and difficult point to some on initial contact with the theory of data; the third section of this chapter is designed to clarify it.

The term data is used here to refer only to that which is analyzed. As will be evident, the same observations may frequently be interpreted as one of two or more different kinds of data. The choice is an optional decision by the scientist and represents a creative step on his part in collecting the data he analyzes. It is the different kinds of data and their interrelations with which this theory is concerned. It might help to clarify the scope of the theory of data if we turn to the diagram of Fig. 1.1.

At the extreme left is the universe of potential observations, all of the things the behavioral scientist might choose to record. If an individual is asked whether he would vote for candidate *A*, the observer usually records his answer, yes or no; but we might ask why the time it took him to answer is not of interest, or whether there was a change in respiration, or in his galvanic skin response, or what he did with his hands, and so on. From this richness the scientist must select some few things to record, and this step is called phase 1 in the diagram.

These recorded observations, however, are not yet data in the sense of this theory of data; an interpretive step on the part of the scientist, called phase 2 in the diagram, is required to convert the recorded observations into data. Phase 2 involves a classification of observations in the sense

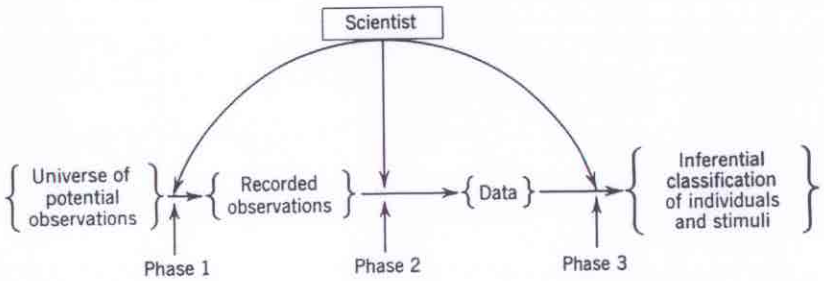

Fig. 1.1 Flow diagram from the real world to inferences.

that individuals and stimuli are identified and labeled, and the observations are classified in terms of a relation of some kind between individuals and stimuli, or perhaps just between stimuli.

Then, given this prior classification, phase 3 involves the detection of relations, order, and structure which follow as a logical consequence of the data and the model used for analysis.

The scientist enters each of these three phases in a creative way in the sense that alternatives are open to him and his decisions will determine in a significant way the results that will be obtained from the analysis. Each successive phase puts more limiting boundaries on what the results might be. At the beginning, before phase 1, there are, perhaps, no limits on the potential conclusions; but each phase then constrains the universe of possible inferences that can ultimately be drawn from the analysis.

For example, suppose an experimenter has subjects judge which of two policies is more beneficial to society. He might use such judgments to study the statements or to study the individuals. If the experimenter decides to interpret the behavior he observed as relations between an individual and a stimulus, he will ultimately come to different conclusions than if he interprets the observed behavior as relations between stimuli. This is an example of a decision which has to be made in phase 2. The decision then restricts the variety of models among which choice must be made in phase 3.

I am not deploring these creative roles of the scientist but merely trying to detail explicitly the processes by which conclusions are drawn about individuals and stimuli from behavioral observations. The basic point is that our conclusions, even at the level of measurement and scaling (which seems such a firm foundation for theory building), are already a consequence of theory. A measurement or scaling model is actually a theory about behavior, admittedly on a miniature level, but nevertheless theory; so while building theory about more complex behavior it behooves us not to neglect the foundations on which the more complex theory rests.

This illustrates the general principle that all knowledge is the result of theory—we buy information with assumptions—"facts" are inferences,* and so also are data and measurements and scales.

The theory of data, then, is concerned with behavioral theory at the initial level that provides the foundation for psychological measurement and scaling. It is concerned only with phase 2 and phase 3, that is, the mapping of the recorded observations into data and the choice of models for making inferences from the data. Phase 1, perhaps the most important of all, the decision as to what to observe, is beyond the scope of this theory.

* In this regard see Hanson (1958), who maintains that even the observations are inferences (or see Putnam's 1959 review of his book).

The kind of analysis in which we are interested has been referred to somewhere as the internal productivity of data—the structure or the relations within the a priori classification system—as distinct from external productivity—the relation of this classification system to that associated with some other set of data.

In still other terms, the entire process of measurement and scaling may be said to be concerned with data reduction. We have, perhaps, an overwhelming number of observations, and their comprehensibility is dependent on their reduction to measurements and scales. This is a mechanical process, but only after buying a (miniature) theory of behavior. It is the universe of these behavior theories, their structure and the relations among them, with which the theory of data is concerned.

Unfortunately, the distinction between the recorded behavior and the data makes exposition difficult. The desirability of illustrations in the course of exposition is obvious. The danger lies in the possibility of inferring that what is done in analyzing a particular behavioral example is in some sense "right" or what should be done. On the contrary, *there is no necessary interpretation of any behavioral example as some particular kind of data.* On the other hand, there are many instances of conventions so nearly universal that an identity appears to exist between the observations and the data. Instances of such conventions abound in the physical sciences and in experimental psychology; for example, when a recorded observation is itself a measure, as the number of drops of saliva, the amplitude of a galvanic skin response, or the number of items right in a test. In such instances the convention seems so natural and reasonable that we are almost unable to make a different interpretation, a different mapping, of the behavior into data.

One of the objectives of this exposition is to loosen these bonds or at least bring about a full awareness of them. To accomplish this, each kind of data will be illustrated by a variety of behavioral examples, following convention in most instances. In other instances the same behavioral examples will be used in less conventional ways to illustrate different kinds of data and the different kinds of results obtained as a consequence. This is done more freely in Parts 2–6.

2. BASIC CONCEPTS

One of the consequences of the formal analysis of what constitutes behavioral data is its classification into four basic kinds of data called *preferential choice* data, *single stimulus* data, *stimulus comparison* data, and *similarities* data. Parts 2 to 5 of the book deal with these in turn and in some detail. The purpose of this section is to introduce on a verbal

and intuitive level the basic concepts of data theory that lead to this classification. This discussion provides the motivation for the formal statement of the theory of data presented in the next section and also serves as an introduction to the next four parts of the book.

To introduce the basic concepts for each of the four kinds of data, we take an imaginary sample of subjects and stimuli and assume that certain observations have been made. The observations we shall consider first are of a kind that maps the most naturally into what we have called preferential choice data. Then, in turn, we consider other kinds of observations that map easily into each of the other three kinds of data. In each instance we map the observations into data by making an interpretation which is extremely common but by no means necessary. In the rest of the book the mapping of observations into data is taken up again in greater detail, and alternatives to the interpretations made here are discussed.

Suppose we have the preferential choices of each of a number of individuals with respect to a number of stimuli. Although it does not matter who the individuals are or what the stimuli are, it is more interesting to imagine that the stimuli are drawn from a reasonably homogeneous class and that the individuals are drawn from a relevant homogeneous class. The stimuli may be candidates, colors, or candies, but not all three or even two. It seems complicated enough at this stage to observe whether an individual prefers a chocolate cream to peanut brittle without asking if he likes the color blue even more. If the stimuli are candidates, the individuals would presumably have some cultural commonality with the candidates. If the stimuli are colors, we might at least want to know if any of the individuals were color blind. If the stimuli are candies, anyone with normal taste perception would do. Finally, for no very important reason, imagine that the stimuli are presented in pairs and the individual indicates his preference in each pair.

We are not surprised to observe that the choices individuals make are not alike. Some of the individuals will yield a set of preferential choices which are transitive, and hence each one's preferences may be completely represented by a rank order of the stimuli from most to least preferred. These rank orders will certainly not all be alike, however. Other individuals will yield preferential choices that are not transitive and so cannot be completely represented by a rank order. What do these individual differences mean? How might they have come about?

We might hypothesize that the stimuli were "really" different stimuli for different individuals. Of course this may not be so from an objective (the experimenter's?) point of view—peanut brittle is peanut brittle, red is red, and candidate A is always candidate A—but subjectively this might

be the case. To one person the only salient characteristic of peanut brittle is that it is brittle; to another the only thing that matters is that it is sticky. Red is red unless the individual is color blind, and candidate *A* is trying to be different things to different people. If we take this point of view, an intuitively reasonable one sometimes, we should not have wasted time collecting preferential choice data. An anchor point is needed, and the same stimulus being presented to different individuals provides such an anchor. If a stimulus differs in a significant way from one individual to the next, absolutely nothing can be done with just these observations, and with this point of view, to try to find out anything about the stimuli or about the individuals.

Having abandoned this hypothesis—that individuals differ in their preferences because they perceive the stimuli differently—we concede that each stimulus is more or less the same thing for everyone, not just in its physical dimensions but in whatever its subjective characteristics might be. Peanut brittle is crunchy, sticky, and sweet; red is—well, red; and candidate *A* is a liberal, internationally minded, inexperienced politician. In fact, we might conveniently imagine that each stimulus can be represented by an appropriately selected point in a space of one, two, three, or more dimensions. At this stage we do not know yet how many dimensions this space should have nor where the points would be that correspond to the stimuli. These, in fact, are questions we would consider asking the data to answer. So the notion of a psychological space with stimuli mapped into points in it is introduced.

If each stimulus point, however, has the same location in this space for all the individuals, how might differences in preferences have arisen? We suspect that the individuals themselves are somehow different, and hence their differences must be captured in the model in order that it may generate their different preferences. We intuitively accept the idea that one individual likes some particular thing more than another individual does, and furthermore, that if this thing were changed somewhat, one individual's preference for it might increase and the other's decrease. It is as if there were, perhaps, an ideal choice for each individual, a stimulus that he would prefer to all the possible alternatives of that kind. We conceive, then, of representing an individual by a point in the same space containing the stimulus points, in such a way that the point corresponding to the individual is a point of his maximum preference in this domain of stimuli.

Consider, for example, a set of statements of opinion about athletic scholarships ranging from for to against. An informed individual will feel that some of these statements reflect his opinion better than others. In fact, we could conceive of some hypothetical statement that just

exactly expresses how he feels about athletic scholarships. This would be a statement he would endorse in preference to all other statements and would correspond to an ideal point.

We immediately leap to the further hypothesis that the individual's preference ordering reflects how near the corresponding stimulus points are to his ideal point. We now have a psychological space with both stimuli and individuals mapped into points in such a way that the mutual relations among the points in the space reflect, by some rule, the observed preference orderings of the various individuals. Now we try to account for individual *differences* in preference orderings in terms of the location of ideal points in a common frame of reference with the stimuli.

This theory of preferential choice behavior leads to an algorithm, called the unfolding technique, for constructing a psychological space from such data. Most of Part 2 is concerned with the theory of this technique. A space such as this in which there are points corresponding to individuals and to stimuli is called a *joint space* in this book.

One further step in the process of abstraction should be taken. The important distinction is not that the points correspond to individuals and to stimuli in themselves, but rather to elements of two different sets of real world objects. For example, if we observed in a concept formation experiment the rank order in which each stimulus matched the several concepts, the data could be analyzed by unfolding to arrive at a joint space of stimuli and concepts.

We have introduced some of the basic concepts associated with the theory of preferential choice, but we considered only a limited portion of our hypothetical observations. Let us return to them and recall that some individuals had pairwise preferences which were intransitive. Such observations make it incumbent upon us to show how the relations among these points in the joint space could somehow generate intransitivities. We may choose to construct the model so that intransitive choices occur only as a consequence of fluctuations. The argument would be that in some fundamental, basic sense individuals' preferences are "really" transitive but that some random component blurs this basic picture. In line with this, it might be imagined that the ideal point corresponding to an individual in the joint space has a certain random oscillation and so also have the points corresponding to the stimuli. In fact, most individuals do not have a sharp point of maximum preference with respect to candy, color, candidates, or anything else. Furthermore, the same physical stimulus appears somewhat differently at different times. We might hope, however, that in relation to the total space these regions of oscillation or uncertainty for a point are reasonably small.

If this were the case, collecting data by the method of pair comparison

would yield intransitive sets of preferential choices not by virtue of any inherent intransitivity but by virtue of a random component. How might we distinguish between such a random component and a more fundamental "true" intransitivity? By repeated independent observations a random component could be controlled enough to reveal the predominant, more usual, and hence characteristic preference. Then the question could be raised whether these stochastic preferences satisfied transitivity or whether there was some significant degree of intransitivity. If intransitivity of the predominant stochastic preferences was obtained, then the unfolding model would be violated and a different theory would be called for—one that can accommodate significant intransitivities.

Since there may be random oscillation of a point there is a need for methods of collecting data that yield what might be called redundancy. The term redundant data is used here to refer to experimentally independent replications of an observation. There exists a systematic way of looking at most methods of collecting data which structures and relates them with respect to their capacity to generate information and with respect to how much of the information is redundant under certain assumptions. Since these matters are relevant and are used throughout the rest of the book, they are taken up in the next chapter. They are not an intrinsic part of this theory of data but a very useful parallel development. They have to do with certain abstract properties of the recorded observations from which the theory of data springs.

Before we leave this discussion of preferential choice data, we summarize what has been abstracted as their essential character. Individuals and stimuli are regarded as points in a psychological space. The preferential choice of an individual between two stimuli is interpreted to mean one stimulus point is nearer the individual's ideal point than is the other stimulus point. The model is saying that there is a distance, as yet undefined, between every pair of points, and in particular, between an individual's ideal point and a stimulus point. Here the data consist of pairs of points, sometimes called *dyads* or *couples*, in which the elements of a pair correspond, in order, to an individual and to a stimulus; such pairs of points are referred to as being from distinct sets. The data consist of more than that, however; they consist of the information that the elements of one pair of such points are nearer each other than are the elements of another pair. So, we might describe such data in general terms as comprising order relations on pairs of dyads whose elements are from distinct sets.

In this particular set of hypothetical data, in which an individual indicates a preference between a pair of stimuli, an important question is whether the point corresponding to the individual is one and the same

point in both dyads, whether the two points might differ by virtue of just a random component, or whether they might be points drawn from entirely different multivariate distributions. For example, if some of the stimuli are candies and some are cigarettes, then an individual judging whether he prefers peanut brittle to Camels may correspond to one ideal point for the candy and to another one for the cigarette. The unfolding theory may be used to test hypotheses of this kind against the data.

We turn now from preferential choice data to single stimulus data. Suppose our hypothetical individuals are presented with the same stimuli as before, but this time each is asked whether he would approve of that candidate, or whether he would like that color for his living room, or whether he would buy that kind of candy. In generic terms the individual merely responds positively or negatively to some degree to each of the stimuli in turn or makes an absolute judgment about each. The previous model was deliberately constructed for preferential choice behavior; the kind of observation being made here is what is called single stimulus observation. How may the previous model for preferential choice be adapted to handle single stimulus data?

The individual is identified with his ideal point in the same space with all the stimulus points. He says "yes" to some and "no" to the others. A reasonable hypothesis is that he likes those stimuli whose points are not too far distant from his ideal point. We might hypothesize that for each individual there is a critical neighborhood at the moment he responds to a particular stimulus which provides that he will respond positively if the stimulus is within it, otherwise not. The essential character of such data may be stated in abstract terms analogous to that for preferential choice data. Individuals and stimuli are mapped into points in a psychological space, and the individual's positive response to a stimulus is taken to signify that the stimulus point is in some sense "near" the ideal point. Here the data consist of pairs of points in which the elements of a pair correspond, in order, to an individual and a stimulus. So the dyads are made up of points from distinct sets. The information in the data indicates whether the distance between a pair of such points is or is not greater than some threshold or critical distance. Such data may be described in abstract terms as a *proximity* relation on a pair of points from distinct sets.

The value of an abstract formulation lies in its generality. A great variety of real world observations may be mapped into the same kind of data. Clinical diagnosis, rating scale behavior, and magnitude estimation (to run some kind of a gamut) are examples. To diagnose is to attach a syndrome label, such as "schizophrenic," to a patient. Supervisors asked to rate the efficiency of foremen associate a descriptive adjective with each foreman. Individuals asked to estimate the weight of a stimulus

associate a number with it. In each case the real world consists of two distinct sets of elements—patients and syndromes, foremen and adjectives, weights and numbers. Each element is assumed to correspond to a point in a space. There is a point corresponding to each syndrome, the textbook description, and the point corresponding to the characteristics of the patient as perceived by the clinician. If the clinician answers "yes" to the question, "Is this patient schizophrenic?," his answer may be interpreted as a proximity relation on the corresponding pair of points. In the case of rating scales and magnitude estimation, the objects of judgment (foremen or weights) and the response categories (adjectives or numbers) are assumed to be points on a line, and again the judgment is interpreted as a proximity relation on a pair of points from distinct sets.

The fact that these very different kinds of observations may be looked on as generating the same kind of data means that the models constructed for the analysis of these several kinds of real world observations are intimately related and, at least potentially, there is a certain interchangeability among them. As we proceed we shall perceive more clearly the similarities and differences among models in terms of the assumptions they make about such things as the information in the data. For some, coarse grain data like "near-far" are sufficient, and for others the interpoint distances must be measured on a ratio scale. Hence we can sometimes observe a hierarchical relation among models in terms of the information they require over and above the elementary proximity relation.

Up to this point two kinds of observations have been illustrated and given a formal interpretation; the first was an instance of preferential choice data, the second an instance of single stimulus data. The formal character of the data in these instances may be seen to differ in two respects. One difference is that a relation exists on a pair of dyads for preferential choice data, whereas it is on a pair of points for single stimulus data. A second difference is that for the preferential choice data it is an order relation, and for single stimulus data it is a proximity relation.

The question which naturally leaps to mind is whether there are observations that might be interpreted as proximity relations on pairs of dyads and whether there are other observations that might be interpreted as order relations on pairs of points. The answer is "yes." We leave these variations for later detailed discussion, however, and pursue instead another basic concept, which has to do with a third dimension in which data may differ.

Returning once more to the candidates, colors, or candies, let us present them in pairs to our subjects and ask each individual to judge which member of each pair is the more liberal candidate, the warmer color, or the more subtle blend of flavors. Usually when we make such

observations we find that, by and large, the various judges tend to agree with each other. There is a significant degree of conformity. In fact, when it is absent, the experimenter tends to separate the subjects into two or more groups, each group containing subjects who are more homogeneous. The objective of the experimenter in making such observations is to "measure" the stimuli on the subjective attribute in question. The individuals are replicated to control a random component in the judgment. The point here is that judgments are presumed to reflect differences among the stimuli, not the individuals. If the individuals are presumed to be different, they are put in separate groups and their judgments analyzed independently.

The elemental observation here is the judgment of an individual that of the two stimuli presented one has more of the specific attribute in question than the other. How might such observations be abstracted? We might hypothesize that each stimulus may be represented by an appropriate point on a line which is a continuum to be interpreted as the attribute in question. Furthermore, when an individual judges one stimulus to have more of this attribute than the other, this corresponds to the statement that the point for that stimulus is to the right of the other on this continuum. The same individual at different times might contradict himself, or different individuals might not always agree, but this is accounted for by the random "oscillations" of each point within a more or less restricted neighborhood.

When we make such an interpretation of the observations, the essential character of the data may be stated in abstract terms similar to those for preferential choice data and for single stimulus data. In this case there are pairs of points in which both points have been drawn from the same set, as both points correspond to stimuli. The individual in such an interpretation is not a point in the space. Finally, the observation has been interpreted as an order relation on that pair of points. To put it succinctly, we have order relations on pairs of points from the same set—data that might be called *stimulus comparison data* in contrast to preferential choice data and to single stimulus data.

Such data introduce a new dimension in which data may differ. We have already seen that data may differ with respect to *whether a relation exists on a pair of points or on a pair of dyads*, and with respect to *whether the relation is an order or a proximity relation.* In the previous examples of preferential choice data and of single stimulus data, the elements of a pair of points were drawn from distinct sets, that is, individuals and stimuli, or objects of judgment and response categories. In the case of stimulus comparison data the elements of a pair of points are seen to be from the same set; the pairs of points all correspond to pairs of stimuli.

This is perhaps one of the oldest kinds of data in the history of experimental psychology, and it is not surprising to find that there are a variety of models for analyzing such data to arrive at measures of the stimuli on the attribute in question. These models differ in ways that are discussed in Part 4.

This overview of the theory of data has now introduced the three fundamental dichotomies in terms of which I propose to categorize all behavioral data. Before turning to more abstract exposition, however, we need to describe one more kind of data in order to complete this survey. In the discussion of preferential choice behavior at the beginning of this section, we considered the hypothesis that individual differences in preferences might have arisen because a stimulus was not subjectively the same thing to all judges. We now turn to this problem of finding out in what terms different individuals perceive a stimulus that to the experimenter is unchanging.

It might seem that stimulus comparison data of the kind just discussed would serve this purpose, because we could "measure" how liberal the candidates are, how warm the colors, and how subtle the blend of flavors in the candies. There is no apparent limit, however, to the variety of attributes on the basis of which we might coerce the subjects into comparing the stimuli. And the question naturally arises: If we permit the subject freedom of choice in evaluating or comparing stimuli, in what terms does he do it?

To see how this type of problem may be approached, let us return to our now overworked subjects and ask them to perform one more task. All pairs of stimuli will be formed, and the individuals will be presented with these pairs and asked in which pair the members are more alike. This, of course, might be done in a great variety of ways; to be concrete let it be done by pair comparisons, that is, pairs of dyads are presented and the individual says which one of the dyads contains the more similar stimuli. How might such behavior be represented in terms of the type of model that we used in discussing the other kinds of data?

The individual might be presumed to perceive each stimulus as a union or coalition of certain characteristics or attributes. Once again the stimulus may be represented by a point in a space in which the coordinates of the point correspond to the projections of the stimulus on the various dimensions (characteristics) which are the relevant ones in the individual's perception of the stimulus. In fact, all the stimuli may be presumed to be represented by points in a common space of relevant dimensions. Note that no constraint has been placed on the individual about which or how many dimensions there might be. In fact the object in making these observations is to attempt to determine just that.

The individual's judgment that one pair of such stimuli are more alike than another may then be interpreted to mean that the distance between the one pair of points is less than that between the other pair of points. In formal terms, such data are order relations on pairs of dyads in which all the points are from the one set of stimulus points. In verbal terms, what the data consist of are comparisons between stimulus differences, and hence such data are called *similarities data.* The individual is not here conceived of as corresponding to a point, but rather as being characterized by the structure of the entire space. Information about this structure is what we hope to extract from the data. A model designed to obtain this needs to make further specifications, such as what is meant by distance in the space. Then, given these assumptions, a calculus that will construct a space which can reproduce the data to some significant degree may be built. Such data and the models designed for their analysis are discussed in Part 5.

Some of the interrelations among these four different kinds of data are discussed in the chapters of Part 6, which concludes with a chapter relating this formulation of a theory of data to an earlier formulation.

3. THE FORMAL BASIS OF DATA

The formal axioms and definitions that comprise this theory of data are contained in an appendix to this chapter. Some models would require more axioms, and some would require less. The compromise I have made is to provide as limited a number as would still have sufficient breadth and scope to span the variety of models with which this book is concerned. In this section we discuss these axioms from a primarily heuristic point of view in order to familiarize the reader with much of the notation used throughout the book.

In all behavioral observations converted into data, something plays the role of an acting, deciding, responding organism and something plays the role of a stimulus. The acting organism may be a worm, an organ, a colony, or a social group. The stimulus varies over a gamut from the highly controlled excitation of an end organ to the stimulus complex an individual responds to in social interaction. It is sometimes said that psychology is concerned with the problem of what is the nature of the stimulus. Data theory does not have this problem, however. Data theory merely says that whenever behavior has been mapped into data to be analyzed, someone has labeled something as a stimulus and something else as a behaving organism. Data theory neither approves nor disapproves of this labeling.

Of course, responses also exist. What characterizes measurement

situations is that the repertoire of responses is strictly limited. Only in certain projective instruments is the effort made to free the response of any restraints. The repertoire of responses which lead to data and measurement has, generically speaking, a very limited variety. The responses reflect either consonance or dominance. This consonance or dominance relation may be between an individual and a stimulus, as when an individual endorses a candidate or solves an arithmetic problem, respectively, or between stimuli, as when he judges them similar or says one is greater than another. The objective of these observations is to measure the individuals and/or the stimuli on the one or more relevant attributes or traits.

These remarks provide some insight into the particular axioms that constitute the basis of this theory of data. It is evident that individuals and stimuli can be thought of as points in a space of one or more dimensions in such a way that the relations among the points will reflect the observed behavior, and hence the space is called a psychological space. When the space is one-dimensional it is frequently called a subjective scale.

The first three axioms and the first definition are concerned with the existence and some of the properties of the space a measurement or scaling model is designed to construct. Axiom 1 postulates a space the dimensions of which are segments of the real line. This axiom is stronger than necessary for some measurement models in that it postulates more than is needed. Some measurement models, for example, lead only to ordinal scales or to ordered metric scales, and some multidimensional scaling models only require spaces whose dimensions are represented by ordinal scales. Rather than have a different axiom in the basis for each level of scaling, however, we postulate as much as is ever necessary and then speak of weaker measurement models as "recovering" this space only at lower levels of scales.

The space is assumed to be a metric space, but the distance function is not specified. Most psychological scaling models assume a multidimensional psychological space to be Euclidean, to incorporate the familiar everyday notion of distance. Some do not even require a metric space, whereas others specify a distance function other than Euclidean, such as the "city block" model in which the distance from one point to another is measured along perpendicular paths and there are no diagonal routes.

With this machinery, then, we have the notion of a space with points in it in relatively fixed locations and having one or more dimensions. Our purpose is to use this space as a representation of the behavior of individuals responding to stimuli. For many kinds of behavior or interpretations of behavior such a representation may not be suitable or other

kinds of representations may be more useful. Such behavior or interpretations of it are not of concern in this book.

We turn, then, to the matter of how to utilize these metric spaces for the representation of behavior in terms sufficiently general to cover the range and variety of models in psychological scaling. For notational purposes we have the following label sets:

$$D = \{1, 2, \ldots, d, \ldots, r\}; \qquad H = \{1, 2, \ldots, h, \ldots, t\};$$
$$I = \{1, 2, \ldots, i, \ldots, m\}; \qquad J = \{1, 2, \ldots, j, k, l, j', \ldots, n\}$$

The set D is used to designate the dimensions of an r-dimensional space. The set H designates trials, or another appropriate temporal variable, as when an individual responds to the same stimulus more than once.

The sets I and J are for the designation of two distinct sets of real world objects. An illustration might be a set of individuals and a set of stimuli, in which case the convention is adopted of using the label set I for the individuals and the label set J for the stimuli. In some instances the stimuli being responded to may be other individuals, as occurs when an individual is asked who in a group influences him most. Here the members of the group serve both as individuals responding to stimuli and as stimuli to the other individuals of the group. If the experimenter, in mapping these observations into data, desires to distinguish between an individual as a respondent and the same individual when being responded to (that is, as a stimulus), the label set I is used for respondents and the label set J for the individuals as response objects (stimuli). If the experimenter decides not to distinguish between these two roles, then the individuals constitute only one set of objects.* When there is only one set of objects either label set may be used, of course, but the convention of using the set J is followed. Finally, one other convention is followed in the use of these label sets I and J. Suppose individuals are identifying silhouettes of aircraft. The silhouettes constitute one set of objects, and the identifying labels, the response alternatives, are another set of objects. The label sets I and J would then be used for these two distinct sets, the label set I for the objects of judgment and the label set J for the response alternatives. This occurs in any instance of rating or absolute judgment in which one set of objects is mapped into another set.

Some very critical distinctions made by the theory of data rest on the composition of pairs of points, so we spell out these distinctions in some detail.

We sometimes have one set of objects and sometimes two sets of objects to be mapped into points in a psychological space in such a way that

* This is an example of the creative role the experimenter plays in making data out of the behavioral observations.

observed relations between the real world objects will be represented by abstract relations among the points in the space. If we have only a single set of objects, then the set of points which correspond to these objects is called the set Q, and the objects are called stimuli. If we wish to talk about a second set of distinct objects in the real world, then we need to refer to the set of points which corresponds to them, so we designate that set of points as the set C. The set of objects corresponding to the points in the set C is not uncommonly a set of individuals.

Data sometimes involve a relation between an individual and a stimulus, a heterogeneous pair, and we refer to the corresponding pair of points as a heterogeneous pair or heterogeneous dyad. Such a pair of points, then, is made up of one point from the set C and one point from the set Q. The set of all such pairs we call A. If we imagine a matrix with elements of the set C designating rows and elements of the set Q designating columns, then each cell of this matrix correspond to a pair (c_i, q_j) and all the cells of this matrix constitute the set A. The set A is known as the Cartesian product of the sets C and Q and may be designated $A = C \times Q$. Individuals passing and failing arithmetic problems or endorsing and rejecting candidates may readily be interpreted as A data.

Data sometimes involve a relation between a pair of stimuli, a homogeneous pair, and so we may refer to the corresponding pair of points as a homogeneous pair or homogeneous dyad. Such a pair of points is made up of two points from the set Q. The set of all such pairs we call B, and it is formed from the Cartesian product of the set Q with itself, that is, $B = Q \times Q$. Judgments as to which candidate is more liberal, which candy is sweeter, and which color is brighter may readily be interpreted as B data.

We also have a need for supersets consisting of pairs of dyads. They are needed for data which involve a relation between two homogeneous pairs of points, that is, all four points are stimuli. The set of all the pairs of homogeneous dyads we designate as $B \times B$ data. Observations to the effect that the "confusions" between one pair of stimuli are greater than the "confusions" between another pair of stimuli may readily be interpreted as $B \times B$ data.

Another superset of interest is that for data involving a relation between two heterogeneous pairs of points, that is, each pair of points is a heterogeneous pair. The set of all the pairs of heterogeneous dyads we designate as $A \times A$ data. Judgments to the effect that Mr. X is a better president than Mr. Y is a governor could be examples of $A \times A$ data. Such data have not been given any serious attention by either collectors or analyzers of data but a proper subset has. This subset, which is of some interest, is the superset of pairs of heterogeneous dyads which have one

point in common, that is, each pair of points is a heterogeneous pair but one member is the same in each pair. Judgments to the effect that Mr. X would make a better governor than president could be an example of such data.

Because an individual's point is typically the common element in a pair of heterogeneous dyads, we designate the set of all dyads involving some particular point in common A_i. The set of all the pairs of heterogeneous dyads involving a particular individual is, then, $A_i \times A_i$; and the set of all the pairs of heterogeneous dyads which have a point in common is the union of the $A_i \times A_i$ for all i, which we designate $\bigcup_i (A_i \times A_i)$.

A possible example of $\bigcup_i (A_i \times A_i)$ data would be the pair comparison preferences of a number of individuals for various kinds of candies. A heterogeneous dyad is made up of an individual's point and a candy point; and the individual's point is common to the pairs of such dyads which are compared.

The important axiom that generates the basic kinds of data is Axiom 4. It states that all psychological data may be viewed as an interpretation of behavior in which three dichotomies are satisfied, as stated in the appendix.

The first two dichotomies generate the various sets which have just been constructed, as is shown in the next section. In effect, what the first two dichotomies assert is that all data may be viewed as a relation on a pair of points or on a pair of dyads. To each dyad corresponds a distance between the member points, and a distance may itself be regarded as a point, so we might say that all data may be viewed as relations on pairs of points; it is convenient, however, to segregate data which consist of relations on distances between points from data which consist of relations on the points themselves, because the models for scaling points may need to be different from those for scaling distances.

The third dichotomy specifies that the relation is either an order relation or a proximity relation. For example, a judgment as to which of two candidates is more liberal may be interpreted as an order relation on the corresponding pair of points. The judgment that a particular candidate is liberal may be interpreted as a proximity relation on the corresponding pair of points—one representing the candidate and the other representing the concept "liberal." As alternatives to these names for the relations the terms *dominance* and *consonance* relation, respectively, might sometimes capture the real world implication more adequately.

In its most summary form, then, the theory of data asserts that behavior may be interpreted as a relation on a pair of points and the relation may be either a dominance or a consonance relation.

We come then to Axiom 5 and the final four definitions. These are all

concerned with the concept of *relevant dimensions*. The point is that although an individual and the stimulus he is responding to have many characteristics, at the moment that he responds only certain of these characteristics may be relevant in mediating that response. Thus, one of the characteristics of a candidate for office is his religious affiliation, and, similarly, the voter also has a religious affiliation. This characteristic or dimension, however, may or may not be relevant in that voter's evaluation. In general we might anticipate that not all the characteristics of an individual and a stimulus would mediate his response to it. It is important to make these considerations explicit because they are implicit in all models from the point of view of this theory of data, in that all models require a unique identification between a set of relevant dimensions and a stimulus when individuals respond to it.

Definitions 2 and 3 designate the projection of a stimulus point and an individual point respectively into the subspace of relevant dimensions. Thus c_{hij}, for example, is the point corresponding to individual i at the moment h when responding to stimulus j. Similarly q_{hij} is the point corresponding to stimulus j at the moment h when individual i is responding to it.

The last two definitions designate the value of the function $p(a, b)$ for a heterogeneous pair and a homogeneous pair, respectively, in the space of relevant dimensions. Thus the quantity $|p_{hij}|$ is the distance between an individual point and a stimulus point when projected into the space of relevant dimensions.

With this set of axioms and definitions, the variety of psychological data that are recognized may now be constructed and related to various psychological measurement and scaling models.

4. THE STRUCTURE OF DATA

Axiom 4 offers three dichotomies and asserts that all behavioral observations may be so interpreted as to satisfy each of these three dichotomies. This suggests the possibility of their being, in principle at least, $2^3 = 8$ different kinds of data. This cube can be oriented in any of three different ways, and it seems to make the most general psychological sense to subordinate the third dichotomy somewhat to the first two. The cross partition of the first two dichotomies, then, yields four classes called quadrants, each of which is dichotomized by the third condition of Axiom 4. The quadrants are numbered from QI to QIV and each is divided into an "a" and a "b" half. Their organization is portrayed in Fig. 1.2.

I suggest that these eight classes represent the eight kinds of primitive

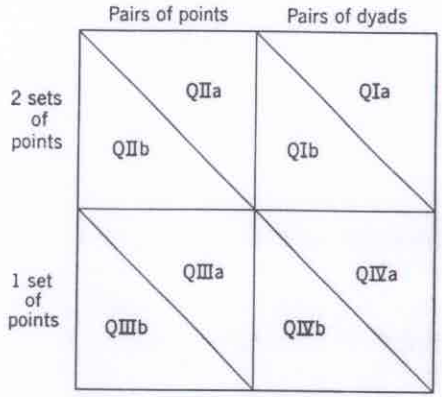

Fig. 1.2 The eight kinds of data.

data from which all psychological measurement arises. A psychological measurement model is designed to take the data of one of these classes and make inferences about a psychological space.

The reason for the subordination of the third dichotomy to the other two lies in the fact that, in some instances, observations are made which, taken together, are a mixture of the two classes of data in a single quadrant. This occurs most commonly when an individual is allowed an intermediate category of response. This point is illustrated in several instances as we once more range over the varieties of data, somewhat as before but more compactly, to illustrate the abstractions involved and to define the information in each class of data.

Quadrant Ia

This class of data consists of order relations on pairs of dyads, the elements of which are from distinct sets. The preferential choices of an individual over a set of alternative stimuli offer a natural illustration. The stimuli and the individuals are mapped into a joint space, and the preferential choice is interpreted as an order relation on the relative distances of the two stimulus points from their respective ideal points.

Formally, the information in the data may be defined as

$$|p_{hij}| - |p_{hik}| \leq 0 \Leftrightarrow j \cdot\rangle k$$

where the symbol $\cdot\rangle$ signifies "preferred to." That is, if and only if, at the moment h, the point corresponding to alternative j is at least as near the ideal point of individual i as the point corresponding to alternative k, the individual says he prefers alternative j to alternative k. The definition

of the information in this class of data readily generalizes to ranking more than two alternatives.

Quadrant Ib

Data of this class, to my knowledge, have never been collected in their own right exclusively, and there are no models for analyzing them. This class is important, however, because it is sometimes obtained in combination with QIa data. QIb data are proximity relations on pairs of dyads, the elements of which are from distinct sets. To obtain such data exclusively, we might present an individual with pairs of stimuli and ask him if he does or does not have a preference—but not what his preference is. His answer, yes or no, could be interpreted to signify that the *difference* between the distance of one stimulus and the distance of the other stimulus, each from their respective ideals, is greater or less than some critical threshold quantity.

It would seem silly in this context to collect such data, because it would be so much more informative, if the subject has a preference, to have him name it—which would be QIa data. QIb data would occur, though, if the subjects were permitted an intermediate category of response, such as indifferent, uncertain, or "can't decide."

The information in QIb data may be defined as

$$\left|\,|p_{hij}| - |p_{hik}|\,\right| \leq \epsilon_{hi,jk} \Leftrightarrow j\,\dot{M}\,k$$

where the symbol $j\,\dot{M}\,k$ signifies "j matches k in preference" and ϵ is a nonnegative real number corresponding to some idiosyncratic difference in distances which bounds the individual's willingness to choose. The information in the data has the interpretation that, if and only if, at the moment h, the distances of the two stimulus points from the ideal points are similar (within $\pm\epsilon$), the individual says he cannot choose between them.

Quadrant I

QIa data require that an individual always indicate his preference, QIb that he never indicate his preference but merely whether or not he has one. Obviously, if sometimes he indicates a preference and sometimes he does not, we have a mixture, which is called QI data. The information in such data may be defined as:

$$|p_{hij}| - |p_{hik}| < -\epsilon_{hi,jk} \Leftrightarrow j \cdot\rangle\, k$$
$$\left|\,|p_{hij}| - |p_{hik}|\,\right| \leq \epsilon_{hi,jk} \Leftrightarrow j\,\dot{M}\,k$$
$$|p_{hij}| - |p_{hik}| > \epsilon_{hi,jk} \Leftrightarrow k \cdot\rangle\, j$$

That is, if and only if, at the moment h, the difference in the distance the two alternative points from the ideal point exceeds some value, individual indicates a preference and the choice reflects the nearer a native.

Quadrant IIa

Formally, the data in this class consist of order relations on pair points from distinct sets, and the information may be defined as

$$p_{hij} \geq 0 \Leftrightarrow i \rangle j$$

where the symbol $\rangle$ signifies "i dominates j," for example, passes, exce or says yes. That is if, and only if, at the moment h, the point co sponding to the individual dominates the point corresponding to stimulus, the individual responds positively to the stimulus. This kin data, a dominance relation between a pair of points from distinct set very common. It may be represented by the conventional interpretat of an individual's passing or failing an arithmetic item as in the w known Guttman scalogram model. The individual may be conceived as being mapped into a point corresponding to a measurement of arithmetic ability, the item as being mapped into a point correspond to its difficulty on the same continuum, and the individual's passing failing as an order relation on the two points.

Quadrant IIb

In Section 2 of this chapter this class of data was illustrated by interpretation given to whether an individual would approve of a particu candidate, whether he would like a particular color for his living room, whether he would buy a particular kind of candy. Formally, such d are proximity relations on pairs of points from distinct sets, and information in such data is defined as

$$|p_{hij}| \leq \epsilon_{hij} \Leftrightarrow i\,\dot{M}\,j$$

That is, if and only if, at the moment h, the absolute distance of the po corresponding to individual i from the point corresponding to stimulu is sufficiently small, the individual responds positively to the stimulus approving, endorsing, agreeing, or saying yes, as the case may be. T quantity ϵ, a nonnegative real number, corresponds here to a limiti difference which bounds the individual's willingness to respond affir atively. Other examples of this class of data have been mentioned in t previous discussion.

Quadrant II

By QII we indicate data which are in part QIIa and in part QIIb. In these data, then, the relation may be either an order or a proximity relation on a pair of points from distinct sets. Such data could be obtained if an individual in responding to a statement of opinion were permitted to say (1) he endorsed it, (2) he did not endorse it, it is too radical, (3) he did not endorse it, it is too conservative. The first response is an instance of a proximity relation, the latter two are instances of order relations.

The information in the data may be defined as

$$p_{hij} > \epsilon_{hij} \Leftrightarrow i \rangle j$$
$$|p_{hij}| \leq \epsilon_{hij} \Leftrightarrow i \dot{M} j$$
$$p_{hij} < -\epsilon_{hij} \Leftrightarrow j \rangle i$$

That is, if and only if, at the moment h, the difference between the two points exceeds a certain quantity, the response indicates one is greater than the other; otherwise the response indicates they are matched.

Quadrant IIIa

Typical of the behavior mapped into the data of QIIIa is the judgment of an individual as to which of two stimuli has more of some attribute. Each stimulus is mapped into a point on a line corresponding to measures of the attribute, and the observation is interpreted as an order relation on this pair of points from the same set.

The information in such data is defined as

$$p_{hi,jk} \geq 0 \Leftrightarrow j \rangle k$$

where $j \rangle k$ signifies the observation "j dominates k." That is, if and only if, at the moment h, for an individual i the point corresponding to stimulus j exceeds the point corresponding to stimulus k, the observation is made that j dominates k.

Examples of the observations mapped into such data are the judgments of individuals as to which of a pair of weights is heavier or which of a pair of statements of opinion is more conservative. The intent of such data is to arrive at a scale of subjective magnitude of the stimuli, called a stimulus scale. The points being measured are all members of one and the same set, and hence are called stimuli. The individuals are not mapped into points on the scale, in contrast to QI and QII.

Quadrant IIIb

The data of QIIIb are proximity relations on pairs of points from the same set. An example would be the behavior of an individual in judging whether pairs of stimuli do or do not match each other.

The formal definition of the information in such data is

$$|p_{hi,jk}| \leq \epsilon_{hi,jk} \Leftrightarrow j \, M \, k$$

where $j \, M \, k$ signifies the observation that "stimulus j matches stimulus k." That is, if and only if, at the moment h, the apparent difference between two stimuli is no more than some prescribed positive quantity, the response indicates that they match each other.

This kind of data is just beginning to be useful in the scaling of distances between pairs of stimulus points. The interpoint distances may then be used to recover the stimulus points themselves.

Quadrant III

As before, the data identified with an entire quadrant is a mixture. In this case the data may be either order or proximity relations on pairs of points from the same set. If in judging which of two stimuli has more of some attribute the individual is permitted an intermediate category of judgment, such as "I don't know," some of the judgments will reflect an order relation and some a proximity relation on a pair of points from the same set.

The information in such data is

$$p_{hi,jk} > \epsilon_{hi,jk} \Leftrightarrow j \rangle k$$
$$|p_{hi,jk}| \leq \epsilon_{hi,kj} \Leftrightarrow j \, M \, k$$
$$p_{hi,jk} < -\epsilon_{hi,jk} \Leftrightarrow k \rangle j$$

That is, if and only if, at the moment h, the difference between two stimuli exceeds a certain minimum, the individual judges one to be greater than the other; otherwise he responds with the intermediate category.

Quadrant IVa

The behavior of individuals when presented with two pairs of stimuli and asked which pair is more alike is representative of behavior typically mapped into QIVa. Each of the stimuli is mapped into a point, and the individual is presumed to be responding to the comparative distances

between pairs of points. The response is interpreted as an order relation on pairs of dyads whose points are all from the same set.

The information in such data may be defined as

$$|p_{hi,jk}| - |p_{hi,j'k'}| \leq 0 \Leftrightarrow (j,k) \langle (j',k')$$

where $(j,k) \langle (j',k')$ signifies the response "the pair (j,k) is more alike than the pair (j',k')." That is, if and only if, at the moment h, the difference between the stimuli j and k appears less than the difference between the pair j' and k', the response is made that the pair (j,k) are more alike.

The recent development of models for the analysis of such data reflects a developing recognition of their importance for the study of the perceptual and cognitive space of an individual. This area, called multidimensional psychological scaling, has a very promising future.

Quadrant IVb

The type of behavior that can be mapped into QIVb is the response of an individual to two pairs of stimuli where the response is interpreted as indicating that one pair is no more alike (or different) than the other pair. Each of the stimuli is presumed to be mapped into a point, and the individual is responding to the differences between pairs of points. The response is interpreted as a proximity relation on a pair of dyads in which all points are from the same set.

The information in such data is

$$\big|\, |p_{hi,jk}| - |p_{hi,j'k'}| \,\big| \leq \epsilon_{hi,jk,j'k'} \Leftrightarrow (j,k)\ M\ (j',k')$$

where $(j,k)\ M\ (j',k')$ signifies a response of the form "the pair of stimuli (j,k) is no more alike or different than the pair (j',k')." That is, if and only if, at the moment h, the difference between the stimuli j and k is within a prescribed positive magnitude of the difference between the pair of stimuli j' and k', the individual indicates that the differences match each other.

The method of bisection for constructing stimulus scales is an example of this kind of data in that the subject adjusts a stimulus to be midway between two others. In effect he manipulates a stimulus until he is unable to distinguish the distance between it and the one above from the distance between it and the one below.

Quadrant IV

If in judging the relative similarity of two pairs of stimuli an individual is permitted to respond by saying that one pair is more alike than another or that he cannot decide, the data could be interpreted as QIV data.

The information in such data may be defined as

$$|p_{hi,jk}| - |p_{hi,j'k'}| < -\epsilon_{hi,jk,j'k'} \Leftrightarrow (j,k) \langle (j',k')$$
$$\big|\, |p_{hi,jk}| - |p_{hi,j'k'}| \,\big| \leq \epsilon_{hi,jk,j'k'} \Leftrightarrow (j,k)\ M\ (j',k')$$
$$|p_{hi,jk}| - |p_{hi,j'k'}| > \epsilon_{hi,jk,j'k'} \Leftrightarrow (j',k') \langle (j,k)$$

That is, if and only if, at the moment h, the difference between one pair of stimuli is sufficiently less than the difference between the other pair, the subject indicates which pair is more alike; otherwise he says their differences match.

5. THE FOUR KINDS OF DATA

Because an intermediate category of judgment has the effect of yielding data some of which fall in the a class of a quadrant and some in the b, the dichotomy of an order versus a proximity relation has been somewhat subordinated to the other two dichotomies. On the level of real world behavior there is a similarity between the behaviors that get mapped into the two classes of a quadrant. Giving a name to each of the quadrants which, at least partially, reflects the kind of real world behavior that is typically mapped into it attempts to capture this similarity. This, of course, is contrary to the earlier distinction insisted on, the distinction between the behavioral observations and the data that are analyzed; hence this step is taken with some trepidation. My experience has been that these real world referents help students in acquiring an understanding of the theory of data. The step is taken, however, with the suggestion that reasoning should proceed from the formal properties of the data rather than from these real world labels. No real world behavior necessarily belongs in any particular quadrant.

In QI the relation observed is on a pair of distances where each distance is the distance between a pair of points from distinct sets. A possible, but by no means universal, source of such data is the preferential choices of an individual over a set of stimuli in which he is comparing the distances of stimulus points from an ideal point. Hence the data may be called individual-stimulus differences comparison or *preferential choice* data.

In QII the relation observed is on a pair of points in which the points are identified with elements from distinct sets. This interpretation may be given to mental test behavior, rating scale behavior, and absolute judgment. We may refer to it as an individual-stimulus comparison or by the familiar term of *single stimulus* data.

It is in this quadrant that the recorded observations and the data *may* merge into one. Ordinary physical measurements are observations which

are QII data; as, for example, the number of drops of saliva produced or the amplitude of the excursion of a recording pen. One set of points corresponds to the real numbers and the other set of points corresponds to the observations. The classification and structure in the data follow directly from those in the real number system, and psychological scaling methods are not required.

In QIII the relation observed is on a pair of points that are from the same set—all stimuli. Such data may be called *stimulus comparison* data.

In QIV the relation observed is on a pair of distances where each distance is between a pair of stimuli. This may be referred to as stimuli-differences comparison or *similarities* data.

These names are portrayed in Fig. 1.3.

These four kinds of data are taken up in turn in the next four parts of the book. In each case an introduction precedes the discussion of models for analyzing that particular kind of data. The introduction will discuss the nature of the data, some of the kinds of behavioral observations that are mapped into that kind of data, and something about the organization of the chapters in that part of the book.

Part 6, the last part, is concerned with interrelations among quadrants and other matters that pertain to the system as a whole.

Single stimulus data or Individual-stimulus comparison or A data	Preferential choice data or Individual-stimulus differences comparison or $A \times A$ data
Stimulus comparison data or B data	Similarities data or Stimuli-differences comparison or $B \times B$ data

Fig. 1.3 The four kinds of data.

6. SUMMARY

This chapter offers an overview of the theory of data and explains the organization of the rest of the book. To introduce the basic concepts of the system a vital distinction is drawn between behavior and data, and the latter term is given a highly restricted meaning. The term data is used here to refer to formal relations on points—because these relations are what are analyzed, not the behavior itself. Although the data are an offspring of behavior, the scientist has a much more intimate and creative relation to the process than that of midwife. Behavior does not yield data by parthenogenesis. The role of the scientist in the process is to choose the genus; the behavior then chooses the species. Behavior never acts or speaks for itself in creating data; it only speaks when spoken to, when asked a question. The experimenter selects the repertoire, a particular alphabet of messages, and then the behavior chooses from these alternatives what to play, what the message is to be. Answers are in terms of the questions asked, and to map behavior into a particular kind of data and to analyze this class of data by a particular model is to ask particular questions.

A circuit, on a verbal and intuitive level, is then made through the varieties of data by means of detailed discussion of various kinds of behavior and how they may be mapped into data. This discussion leads up to the abstraction of the universal characteristics of data which form the basis of the theory of data. The axioms and definitions, contained in the appendix to this chapter, are discussed heuristically in some detail. These lead naturally to the organization or structuring of the varieties of data in the form of the four quadrants of a fourfold table, each of the quadrants being further divided in two. The four quadrants represent the four fundamental, qualitatively distinct, kinds of data and, by indirection, four basic kinds of behavior. Their real world characteristics are, somewhat hazardously, captured in the labels assigned to them. From QI to QIV in turn, they are preferential choice data, single stimulus data, stimulus comparison data, and similarities data. The information in each of these kinds of data is formally defined.

7. APPENDIX

The following are label sets:

$$D = \{1, 2, \ldots, d, \ldots, r\}; \qquad H = \{1, 2, \ldots, h, \ldots, t\};$$
$$I = \{1, 2, \ldots, i, \ldots, m\}; \qquad J = \{1, 2, \ldots, j, k, l, j', \ldots, n\}$$

Axiom 1: To each element d in D there corresponds a segment of the real line $K^{(d)}$.

Definition 1: Let $K = \{x \mid x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)}, \ldots, x^{(r)})\}$, where $x^{(d)}$ is in $K^{(d)}$, in which the elements x are vectors in r-dimensional space.

We have the following sets of points in K:

$$C \subset K \quad \text{in which} \quad C = \{c_i \mid i \text{ in } I\}$$
$$Q \subset K \quad \text{in which} \quad Q = \{q_j \mid j \text{ in } J\}$$

so the points in C and Q correspond respectively to the elements of the label sets I and J.

We construct the following dyads:

$$A = C \times Q \qquad A = \{(c_i, q_j)\}$$
$$A_i \subset A, \qquad A_i = \{(c_i, q_j) \mid i \text{ fixed}\}$$
$$B = Q \times Q \qquad B = \{(q_j, q_k)\}$$

where A is the set of heterogeneous pairs and B is the set of homogeneous pairs.

We construct the following sets of pairs of dyads: $A \times A$, $\bigcup_i (A_i \times A_i)$, and $B \times B$.

The set $A \times A$ is the set of pairs of heterogeneous dyads, and the set $\bigcup_i (A_i \times A_i)$ is a subset of $A \times A$ consisting of the pairs of heterogeneous dyads that have a point in common. The set $B \times B$ is the set of homogeneous pairs of dyads.

Axiom 2: There exists a real-valued function $p(a, b)$ defined for a and b in K, such that $|p(a, b)|$ satisfies

$$|p(a, b)| = |p(b, a)|$$
$$p(a, b) = 0 \Leftrightarrow a = b$$
$$|p(a, b)| \leq |p(a, c)| + |p(b, c)|$$

where $\Leftrightarrow$ signifies "implies and is implied by."

Axiom 3: Given two vectors differing only in one component, the sign of p is determined by that one component.

Axiom 4: All psychological data may be viewed as an interpretation of behavior in which

i. a relation exists on a pair of points (a dyad) or on a pair of dyads;
ii. the elements of a pair of points are drawn from two distinct sets or from one set; and
iii. the relation is either an order relation ($>$) or a proximity relation ($\bigcirc$).

Axiom 5: To each triple (h, i, j) and to each quadruple (hi, jk) there corresponds a subset $D' = D'(h, i, j)$ or $D'' = D''(hi, jk)$ of D, that is $D' \subset D$, $D'' \subset D$. The subset D' or D'', as the case may be, is called the *set of relevant dimensions.*

Definition 2: q_{hij} is the projection of the vector q_j in the set of relevant dimensions, D' or D'', as the case may be.

Definition 3: c_{hij} is the projection of the vector c_i in the set of relevant dimensions D'.

Definition 4: $p_{hij} = p(c_{hij}, q_{hij})$ is the value of $p(a, b)$ for the ordered pair (c_{hij}, q_{hij}), which will play the role of a "signed distance" between the pair of points in the space of the relevant dimensions D'.

Definition 5: $p_{hi,jk} = p(q_{hij}, q_{hik})$ is the value of $p(a, b)$ for the ordered pair (q_{hij}, q_{hik}), which will play the role of a "signed distance" between the pair of points in the space of relevant dimensions D''.

REFERENCES

Hanson, N. R., 1958, *Patterns of discovery*, Cambridge University Press, Cambridge, England.

Putnam, Hilary, 1959, Review of Hanson's Patterns of discovery, *Science*, **129**, 1666–67.