# A BASIS FOR SCALING QUALITATIVE DATA*

LOUIS GUTTMAN
*Cornell University*

## 1. INTRODUCTION

IN A GREAT deal of research in the social and psychological sciences, interest lies in certain large classes of qualitative observations. For example, research in marriage is concerned with a class of qualitative behavior called marital adjustment, which includes an indefinitely large number of interactions between husband and wife. Public opinion research is concerned with large classes of behavior like expressions of opinion by Americans about the fighting ability of the British. Educational psychology deals with large classes of behavior like achievement tests.

It is often desired in such areas to be able to summarize data by saying, for example, that one marital couple is better adjusted than another marital couple, or that one person has a better opinion of the British than has another person, or that one student has a greater knowledge of arithmetic than has another student. There has been considerable discussion concerning the utility of such orderings of persons. It is not our intention in this paper to review such discussions, but instead to present a rather new approach to the problem which seems to afford an adequate basis for quantifying qualitative data.

This approach has been used successfully for the past year or so in investigating morale and other problems in the United States Army by the Research Branch of the Morale Services Division of the Army Service Forces. While this approach to quantification leads to some interesting mathematics, no knowledge of this mathematics is required in actually analyzing data. Simple routines have been established which require no knowledge of statistics, which take less time than the various manipulations now used by various investigators (such as critical ratios, biserial correlations, factor analysis, etc.), and which give a complete picture of the data not afforded by these other techniques. The word "picture" might be interpreted here literally, for the results of the analysis are presented and easily assimilated in the form of a "scalogram," which at a glance gives the configuration of the qualitative data.

Description of the practical procedures, as well as the mathematical analysis, must be postponed to other papers. The present paper is devoted to a non-technical discussion of what we mean by a scale.

## 2. THE NOTIONS OF VARIABLE, FUNCTION, AND SIMPLE FUNCTION

First, a word about what is meant by a variable, whether qualitative or quantitative. We use the term in its conventional logical or mathematical sense, as denoting a set of values. These values may be numerical (quantitative) or non-numerical (qualitative).[1] We shall use the term "attribute"

* Presented to the Thirty-eighth Annual Meeting of the American Sociological Society, New York December 4, 1943.

[1] In conventional courses in undergraduate college mathematics it is not ordinarily pointed out that a great deal of mathematics deals with purely qualitative variables. Notions of metrics and quantitative variables can be arrived at by sequences of qualitative classifications. In fact, this is the manner in which our approach to scaling derives a scale ordering.

The reader who is interested might look at a recent departure in textbooks for an introductory course in college mathematics (M. Richardson, *Fundamentals of Mathematics,* Macmillan, 1941). This book gives a simple, entertaining, and mature introduction to the foundations of mathematics. Its emphasis is on understanding, rather than on manipulation. It covers fundamental topics like point sets, the concept of number, and others that are rarely mentioned in ordinary undergraduate curricula and yet are mainstays of mathematical theory. It is only that most of us have been exposed exclusively to certain algebraic manipulations that we conceive such manipulations to be the essence of mathematics. A more sophisticated view is to regard mathematics as unveiling necessary relationships that arise from classifications. Much useless discusinterchangeably with "qualitative variable." The values of an attribute (or of a quantitative variable, too, for that matter) may be called its *subcategories*, or simply *categories*.

An example of an attribute is religion. A person may have the value "Catholic," "Buddhist," "Jewish," "Mormon," "atheist," or some other value of this variable. There is no particular intrinsic ordering among these values. Another example is expression of an opinion. A person may say, "I like the British," "I don't like the British," or "I don't know whether or not I like the British." Another example is, a person may be observed to smile at another person upon meeting him, or he may be observed not to smile.

Quantitative variables are readily recognized and need no discussion here.

A variable $y$ is said to be a single-valued function of a variable $x$ if to each value of $x$ there corresponds a single value of $y$. Thus, if $y$ has the distinct values $y_1, y_2, \ldots, y_m$, and if $x$ has the distinct values, $x_1, x_2, \ldots, x_n$, where $m$ and $n$ may be different, $y$ is called a single-valued function of $x$ if a table of correspondence can be set up like, for example, the following:

| $x$ | $x_1$ | $x_2$ | $x_3$ | $\cdots$ | $x_n$ |
|---|---|---|---|---|---|
| $y$ | $y_3$ | $y_5$ | $y_{m-2}$ | $\cdots$ | $y_2$ |

For each value of $x$ there is one and only one value of $y$. (The converse need not hold: for the same value of $y$ there may be two or more values of $x$.) Obviously, if $y$ is to be a single-valued function of $x$, then we must have $m \leqq n$.

In particular, suppose $y$ is an attribute, say like the above attribute about expression of liking for the British. Then $m = 3$, and we may denote by $y_1$ the statement, "I like the British"; by $y_2$, the statement, "I don't like the British"; and by $y_3$, "I don't know whether or not I like the British." If $x$ is a quantitative variable which takes on more than $m$ values $(n > m)$, and if we can divide the $x$ values into $m$ intervals which will have a one-to-one correspondence with the values of $y$, then we shall say the attribute $y$ is a *simple* function of $x$. For example, suppose $x$ takes on the ten values 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Then the correspondence table might be as follows:

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | $y_1$ | $y_1$ | $y_1$ | $y_3$ | $y_3$ | $y_2$ | $y_2$ | $y_2$ | $y_2$ | $y_2$ |

Or we might show this graphically by plotting the $x$ values on a straight line, and cutting it into intervals:

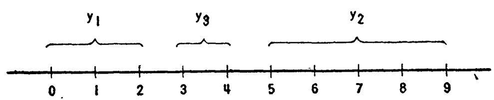


FIGURE 1

For statistical variables, another representation is in terms of a bar chart of frequencies, and this is what we use for convenience in §10 and §11 below.

### 3. THE DEFINITION OF SCALE

For a given population of objects, the multivariate frequency distribution of a universe of attributes will be called a *scale* if it is possible to derive from the distribution a quantitative variable with which to characterize the objects such that each attribute is a simple function of that quantitative variable. Such a quantitative variable is called a scale variable.

Perfect scales are not to be expected in practice. The deviation from perfection is measured by a *coefficient of reproducibility*, which is simply the empirical relative frequency with which the values of the attributes do correspond to the proper intervals of a quantitative variable. In practice, 85 percent perfect scales or better have been used as efficient approximations to perfect scales.

A value of a scale variable will be called a *scale score*, or simply a *score*. The ordering of objects according to the numerical order of their scale scores will be called their scale order.

Obviously, any quantitative variable that is an increasing (or decreasing) function of a scale variable is also a scale variable. For

sion of mathematics as a "tool" in social research could be saved by recognition of the fact that qualitative classifications lead to just as rigorous implications as do quantitative.

example, in the illustration in §2, consider $x$ to be a scale variable. Any constant could be subtracted from or added to each of the $x$ scores, and $y$ would remain a simple function of the transformed $x$. Thus, the scores 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 could be replaced by the respective scores $-5$, $-4$, $-3$, $-2$, $-1$, 0, 1, 2, 3, 4. Or the $x$ scores could be multiplied by any constant, or their square roots or logarithms could be taken—any transformation, continuous or discontinuous, could be used, as long as the rank order correlation between the original $x$ and the transformed variable remained perfect. All such transformations will yield scale variables, each of which is equally good at reproducing the attributes.

Therefore, the problem of metric is of no particular importance here for scaling. For certain problems like predicting outside variables from the universe of attributes, it may be convenient to adopt a particular metric like a least squares metric, which has convenient properties for helping analyze multiple correlations. The interesting mathematics involved here will be discussed in another paper. However, it must be stressed that such a choice of metric is a matter of convenience; any metric will predict an outside variable as accurately as will any other.

In practice, the rank order has been used as a scale variable. (It is in fact a least squares metric for a rectangular distribution of scale scores.)

## 4. THE UNIVERSE OF ATTRIBUTES[2]

A basic concept of the theory of scales is that of the universe of attributes. In social research, a universe is usually a large class of behavior such as described in the introduction above. The universe is the concept whose scalability is being investigated, like marital adjustment, opinion of British fighting ability, knowledge of arithmetic, etc. The universe consists of all the attributes that define the concept. Another way of describing the universe is to say it consists of all the attributes of interest to the investigation which have a common content, so that they are classified under a single heading which indicates that content.

For ease in focusing, let us take an example from opinion research where it is desired to observe the population of individuals in a standardized manner by a checklist of questions. The behavior of interest to the investigation is responses of individuals to such questions. Suppose the universe of attributes consists of all possible questions which could be asked in such a list concerning the fighting ability of the British. Such questions might be: "Do you think the British Army is as tough as the German Army?"; "Do you think the R.A.F. is superior to the Luftwaffe?"; etc. (We do not pause here for problems of wording, interpretation, and the like. The reader is urged rather to focus on the general outline we are trying to establish.) There may be an indefinitely large number of such questions which belong in the universe; and in a particular investigation, ordinarily only a sample of the universe is used.

An attribute belongs to the universe by virtue of its content. The investigator indicates the content of interest by the title he chooses for the universe, and all attributes with that content belong in the universe. There will, of course, arise borderline cases in practice where it will be hard to decide whether or not an item belongs in the universe. The evaluation of the content thus far remains a matter that may be decided by consensus of judges or by some other means. This has been recognized before, although it need not be regarded as a "sin against the Holy Ghost of pure operationalism."[3] It may well be that the formal analysis for scalability may help clarify uncertain areas of content. However, we have found it most useful at present to utilize informal experience and consensus to the fullest extent in defining the universe.

[2] The words *population* and *universe* are ordinarily used interchangeably in statistical literature. For scales, it is necessary to refer both to a complete set of objects and to a complete set of attributes, so it will be convenient to reserve *population* for the former, and *universe* for the latter. In social research, the objects are usually people, so that *population* is appropriate for them.

[3] Clifford Kirkpatrick, "A Methodological Analysis of Feminism in Relation to Marital Adjustment," *American Sociological Review*, June 1939, 4: 325-334.

An important emphasis of our present approach is that a criterion for an attribute to belong in the universe is *not* the magnitude of the correlations of that item with other attributes known to belong in the universe. It will be seen (in §10 below) that attributes of the same type of content may have any size of intercorrelations, varying from practically zero to unity.[4]

### 5. THE POPULATION OF OBJECTS

Defining the universe of attributes is a problem similar to the standard problem of defining the population of objects or individuals[5] of interest to the investigation. An investigator must always delimit the population with which he is working. For example, in the case of opinion about the British as fighters, he must decide *whose* opinions he wishes to ascertain. Is he interested in everyone in the world, or just in everyone in the United States? Is he interested in everyone in the United States, or just in adults? If just in adults, how is an adult to be defined? Here, too, decisions will sometimes be difficult as to whether a particular individual belongs in a population or not, and decisions must be made somehow before the investigation begins, else the investigator will not know whom to observe.

### 6. METHODS OF OBSERVATION

Let us assume that somehow we have a universe of attributes and a population of individuals defined. Next, observations are made as to the behavior of the population with respect to the universe. (In practice this will often be done only with samples. A sample of individuals from the population will have their behavior observed on a sample of attributes from the universe.) How the observations are to be made is of no concern to us here. In opinion research and other fields, questionnaires and schedules have been used. But any technique of observation which yields the data of interest to the investigation may be used. Such techniques for the social and psychological sciences might be case histories, interviews, introspection, and any other technique from which observations may be recorded. The important thing is not how the observations were obtained, but that the observations be of central interest to the investigation.

Use of a questionnaire implies that the investigator is interested in a certain type of universe of verbal behavior. Participant observation may imply that the investigator is interested in a certain type of universe of non-verbal behavior. Such distinct universes may each be investigated separately. It may often be of interest to see how well one universe correlates with another, but such a correlation cannot be investigated until each universe is defined and observed in its own right.

The examples of scales to be given later in this paper happen to comprise observations made by means of questionnaires. It should not be inferred, however, that scaling refers only to that technique. *Scaling analysis is a formal analysis, and hence applies to any universe of qualitative data of any science, obtained by any manner of observation.*

### 7. THE PURPOSE OF SCALING

Obviously it is very clumsy to record the large number of observations ordinarily involved in a universe of attributes for a population of individuals. The recording requires a table with one row for each individual and one column for each attribute. (The table may theoretically be indefinitely large.) It would be convenient if we could represent the observations in a more compact manner which would enable us to reproduce such a table whenever desired. A compact representation, if it could be obtained, would have two great advantages: first, a mnemonic advantage, for a compact representation would be easier to remember than would be a large table; and second, if it were desired to relate the universe to other variables it would be easier to do so by means of the compact representation than by using the large multivariate distribution of the attributes in the universe. From these are derived other ad-

[4] That correlations are no criterion for content has been quite well known. See, for example, R. F. Sletto, *Construction of Scales by the Criterion of Internal Consistency*, Sociological Press, Hanover, N.H., 1937.

[5] For convenience, since the examples in this paper concern populations of human beings, we shall talk entirely in terms of such populations.

vantages which will become apparent as the reader's familiarity with scales grows.

A particularly simple representation of the data would be to assign to each individual a numerical value and to each category of each attribute a numerical value such that, given the value of the individual and the values of the categories of an attribute, we could reproduce the observations of the individual on the attribute. This will be possible only for restricted types of data, where each attribute in the universe can be expressed as a simple function of the same quantitative variable, that is, where the universe of attributes forms a *scale* for the population of individuals.

## 8. AN EXAMPLE OF A DICHOTOMOUS SCALE

As may be expected, the universe of attributes must form a rather specialized configuration for the population of individuals if it is to be scalable. Before describing a more general case, let us give a little example. (A sociological interpretation of this apparently mathematical example is given in §15 below.) Consider a mathematics test composed of the following problems:

(a) If $r$ is the radius of a circle, then what is its area?

(b) What are the values of $x$ satisfying the equation

$$ax^2 + bx + c = 0?$$

(c) What is $de^x/dx$?

If this test were given to the population of members of the American Sociological Society, we would perhaps find it to form a scale for that population. The responses to each of these questions might be reported as a dichotomy, right or wrong. There are $2 \times 2 \times 2 = 8$ possible types for three dichotomies. Actually, for this population of sociologists we would probably find only four of the possible types occurring. There would be the type which would get all three questions right, the type which would get the first and second questions right, the type which would get only the first question right, and the type which would get none of the questions right. Let us assume that this is what would actually happen. That is, we shall assume the other four types, such as the type getting the first and the third questions right but the second question wrong, would not occur. In such a case, it is possible to assign to the population a set of numerical values like 3, 2, 1, 0. Each member of the population will have one of these values assigned to him. This numerical value will be called the person's score. From a person's score we would then know precisely to which problems he knows the answers and to which he does not know the answer. Thus a score of 2 does not mean simply that the person got two questions right, but that he got two particular questions right, namely, the first and second. A person's behavior on the problems is reproducible from his score. More specifically, each question is a *simple function* of the score, as is shown in §10 below.

## 9. THE MEANING OF "MORE" AND "LESS"

Notice that there is a very definite meaning to saying that one person knows more mathematics than another with respect to this sample. For example, a score of 3 means more than a score of 2 because the person with a score of 3 knows everything a person with a score of 2 does, and more.

There is also a definite meaning to saying that getting a question right indicates more knowledge than getting the same question wrong, the importance of which may not be too obvious. People who get a question right all have higher scale scores than do people who get the question wrong. As a matter of fact, we need no knowledge of which is a right answer and which is a wrong answer beforehand to establish a proper order among the individuals. For convenience, suppose the questions were given in a "true-false" form,[6] with suggested answers $2\Pi r$, $(-b \pm \sqrt{b^2 - 4ac})/2a$, and $xe^{x-1}$ for the respective questions. Each person records either a T or an F after each question, according as he believes the suggested answers to be true or false. If the responses of the population form a scale, then we do not have to know which are the correct answers in order to rank the respondents (only we will

[6] We shall assume that no one gets an answer right by guessing. In a later paper it will be shown how scale analysis can actually pick out responses that were correct merely by guessing. But for this, much more than three items are necessary.

not know whether we are ranking them from high to low or from low to high). By the scale analysis, which essentially is based on sorting out the joint occurrences of the three items simultaneously, we would find only 4 types of persons occurring. One type would be $F_1T_2F_3$, where the subscripts indicate the questions; that is, this type says F to question 1, T to question 2, and F to question 3. The other three types would be $F_1T_2T_3$, $F_1F_2T_3$, and $T_1F_2T_3$. These types could be shown in a chart (a "scalogram") where there is one row for each type of person and one column for each category of each attribute. Without going into details, the scale analysis would establish an order among the rows and among the columns which would finally look like this:

| $F_3$ | $T_2$ | $F_1$ | $T_3$ | $F_2$ | $T_1$ |
|---|---|---|---|---|---|
| √ | √ | √ | | | |
| | √ | √ | √ | | |
| | | √ | √ | √ | |
| | | | √ | √ | √ |

FIGURE 2

Or, alternatively, both rows and columns might be completely reversed in order. Each response to a question is indicated by a check mark. Each row has three checkmarks because each question is answered, either correctly or incorrectly. The "parallelogram" pattern in the chart[7] is necessary and sufficient for a set of *dichotomous* attributes to be expressible as simple functions of a single quantitative variable.

From this chart we can deduce that $F_1$, $T_2$, and $F_3$ are all correct answers, or are all incorrect answers. That is, if we were now told that $F_1$ is a correct answer, we would immediately know that $T_2$ and $F_3$ are also correct answers. This means that we can order the men according to their knowledge even if we do not know which are the correct answers and which are the incorrect answers, only we do not know whether we are ordering them from highest to lowest or from lowest to highest. Except for direction, the ordering is a purely formal consequence of the configuration of the behavior of the population with respect to the items. The importance of this fact will become more apparent in more complicated cases where the attributes are not dichotomous but have more than two categories. We do not take the space here to expand on this point, but merely state that the scale analysis automatically decides, for example, where an "undecided" response to a public opinion poll questionnaire belongs, whether it is above "yes," below "no," in between, equivalent to "yes," or equivalent to "no."

### 10. THE BAR CHART REPRESENTATION

Another way of picturing the dichotomous scale of the sample of three items would be as follows: suppose that 80 percent of the population got the first question right, 40 percent got the second question right, and 10 percent got the third question right. The univariate distributions of the three respective items could be shown by the bar chart in Figure 3.

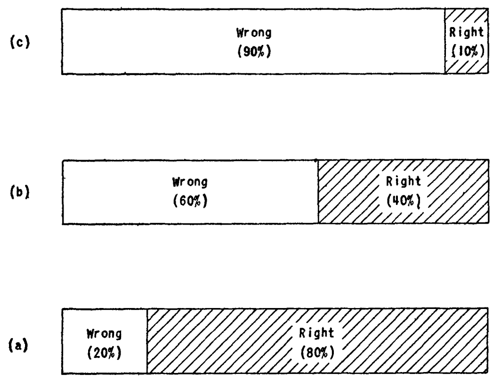


FIGURE 3

The bars show the percentage distributions for the respective questions. The multivariate distribution for the three questions, *given that they form a scale for the population,* can also be indicated on the same chart, since all those who are included in the group getting a harder question right are also included in the group getting an easier question right. Thus, we could draw the bar chart over again, but connect the bars with dashed lines in the fashion shown in Figure 4.

[7] Such a chart, where one column is used for each *category* of each attribute, we call a *scalogram.* The scalogram boards used in practical procedures are simply devices for shifting rows and columns to find a scale pattern if it exists.

Here we can see how the three questions are simple functions of the scores. From the marginal frequencies of the separate items, *together with the fact that the items form a scale,* we are enabled to deduce that 10 percent of the people got a score of 3. The 10 percent who got the hardest question right are included in those who got the easier questions right. This is indicated by the dashed line on the right, between the scores 2 and 3, which carries the same 10 percent of the people (those with a score of 3) through the three bars. The 40 percent who got the second question right include the 10 percent who got the hardest question right and 30 percent out of those who got the hardest question wrong, but all 40 percent got the easiest question right. This leaves us 30 percent who got just the first and second questions right. And so on. Thus we can

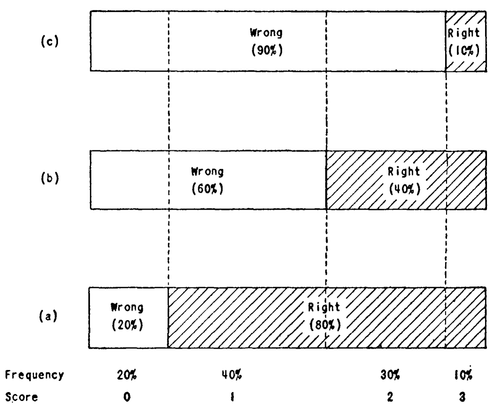


FIGURE 4

think of an ordering of the persons along a horizontal axis, and each item can be thought of as a *cut* on that axis. All those above the cutting point get the question right and all those below the cutting point get the question wrong. Thus there is a one-to-one correspondence between the categories of an item and segments of the axis. Or we can say that each attribute is a simple function of the rank order along the axis.

It is because all the items in the sample can be expressed as simple functions of the same ordering of persons that they form a scale. Each item is perfectly correlated with or reproducible from the ordering along the axis. However, the point correlations between the items are not at all perfect. For example, the four-fold table between the second and third items is as follows:

| | | Question (b) Right | Question (b) Wrong | |
|---|---|---|---|---|
| Question (c) | Right | 10 | 0 | 10 |
| | Wrong | 30 | 60 | 90 |
| | | 40 | 60 | 100 |

The point correlation between the two items is .41. As a matter of fact, the point correlation between two dichotomous items may be anything from practically zero to unity, and yet they may both be perfect functions of the same quantitative variable. That this may be paradoxical might be explained by inadequate treatment of qualitative variables in conventional courses and textbooks on statistics.[8]

An important feature of this four-fold table is the zero frequency in the upper right-hand corner cell. Nobody who got the third question right got the second question wrong. Such a zero cell must always occur in a four-fold table between two dichotomous items which are simple functions of the same quantitative variable.

[8] *Technical Footnote.* A tetrachoric coefficient for the four-fold table above, assuming a bivariate normal distribution, would be unity. However, this is *not* the correlation between the items. It does not tell how well one can predict one item from the other. The tetrachoric coefficient expresses instead the correlation between two quantitative variables of which the items are functions, provided the assumptions of normality are true. The reason the tetrachoric is unity in this case is that the quantitative variables of which the items are functions are one and the same variable, namely, the scale variable. Notice, however, that the distribution of the scale variable according to the rank order is not at all normal. One of the contributions of scaling theory is to do away with untested and unnecessary hypotheses about normal distributions. It is the point correlation that is involved in the mathematical analysis of scaling, not the tetrachoric.

### II. ANOTHER EXAMPLE OF A SCALE

Now let us give an example of a more complicated scale. Suppose we were interested in finding out how much desire soldiers may express now about going back to school after the war is over. Suppose that out of the universe of attributes which define this desire we select the following sample of four questions to be presented on a questionnaire.

1. If you were offered a good job, what would you do?
   (a) I would take the job
   (b) I would turn it down if the government would help me go to school
   (c) I would turn it down and go back to school regardless
2. If you were offered some kind of job, but not a good one, what would you do?
   (a) I would take the job
   (b) I would turn it down if the government would help me go to school
   (c) I would turn it down and go back to school regardless
3. If you could get no job at all, what would you do?
   (a) I would not go back to school
   (b) If the government would aid me, I would go back to school
   (c) I would go back to school even without government aid
4. If you could do what you like after the war is over, would you go back to school?
   (a) Yes
   (b) No

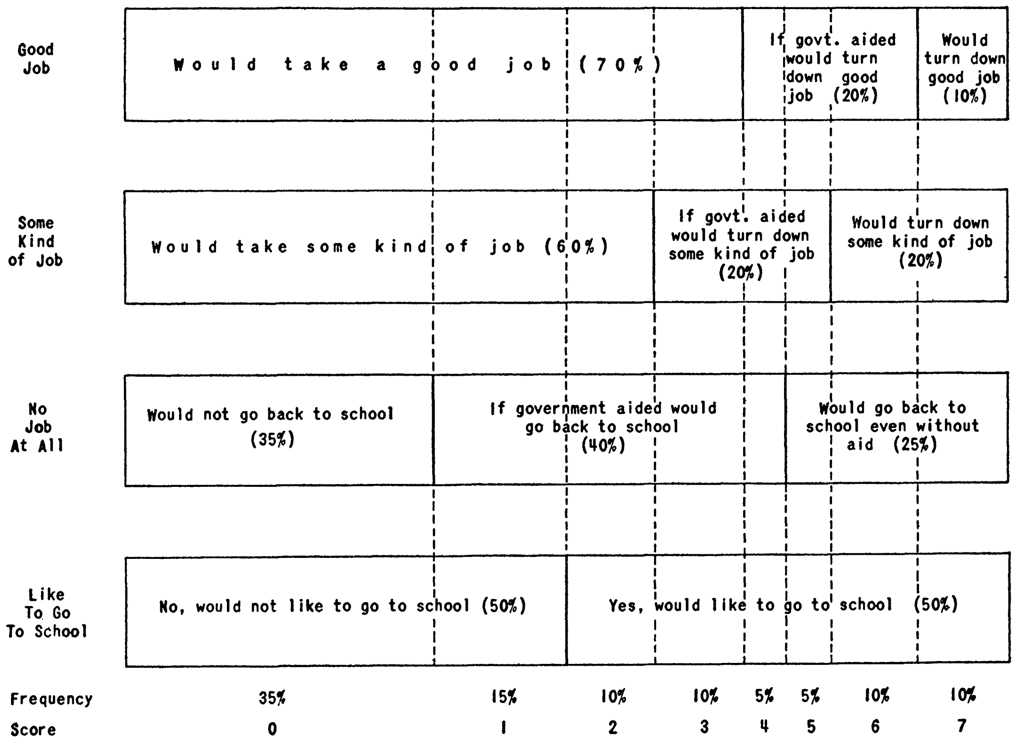


FIGURE 5

Let us suppose the responses of the men to these questions form a scale in the manner shown in Figure 5.

We now know how to read such a chart. 10 percent of the men said they would turn down a good job to go back to school; 20 percent said they would turn down a good job only if the government aided them; 70 percent said they would take a good job; and so on. The 10 percent who said they would turn down a good job are included in the 20 percent who said they would turn down some kind of a job, and the 20 percent are included in the 25 percent who said they would go back to school if they got no job at all, and these 25 percent are included in the 50 percent who said they would like to go back to school.

For three trichotomous and one dichotomous questions there are $3 \times 3 \times 3 \times 2 = 54$ possible types. In order for these to form a scale, it can be shown that at most eight types can occur. The chart shows the eight types, which have been scored from 0 through 7. The chart shows the characteristics of each type. For example, the type with the score 3 includes all men with the following four values: they say that they would take a good job if it were offered to them rather than go back to school; that they would turn down some kind of job if the government would aid them to go back to school; that they would go back to school if the government would aid them if they could get no job at all; and that they would like to go back to school. Thus, by reading the categories crossed by the dashed lines which enclose each type, we can read off the characteristics of the type.

Notice that each of the four attributes is a simple function of the scale scores. For example, the "good job" question has its categories correspond with the following three intervals of the scale scores: 0-3, 4-6, 7.

The question might be raised as to how often will scales be found in practice. Isn't even a fair approximation to a structure like that in the above chart too much to hope for to be found in real life? Towards an answer to this, we can only cite thus far our experience with research in the Army. Literally dozens of sufficiently perfect scales have been found in various areas of attitude, opinion, and knowledge. The example given above of desire to go to school is a fictitious version of a set of similar questions that have actually proved scalable for the Army. Many varieties of data have been found scalable, and many have not. Those data which proved scalable could then be related to other variables very easily. Those that were not scalable required a more complicated analysis to handle them properly.

### 12. ON SAMPLING THE UNIVERSE OF ATTRIBUTES

An important property of a scalable universe is that the ordering of persons based on a sample of items will be essentially that based on the universe. If the universe is a scale, what the addition of further items would do would be merely to break up each type given by the sample into more differentiated types. But it would not interchange the order of the types already in the sample. For example, in Figure 5 above, type 6 would always have a higher rank order than type 5. People in type 6 might be ordered within the type into more subcategories; people within type 5 might be ordered into more subcategories; but all subcategories within 6 would remain of higher rank than all those in type 5. This may be seen in reverse, for example, by deleting one of the questions and noticing that all that is accomplished is to collapse the number of types to a smaller number so that two neighboring types may now become indistinguishable; but any types two steps apart would still remain in the same order with respect to each other.

Hence, we are assured that if a person ranks higher than another person in a sample of items, he will rank higher in the universe of items. This is an important property of scales, that *from a sample of attributes we can draw inferences about the universe of attributes.*

One of the criteria for selecting a sample of items is to choose a sample with enough categories to provide a desired amount of differentiation between individuals. Thus if individuals are desired to be differentiated say only into 10 groups, items should be chosen which will yield 10 types.[9] The shape of the distribution of the rank orders in a sample of attributes will of course depend upon the sample. One sample of attributes may give one shape distribution; another sample may give another shape distribution. This need not be a matter of concern, since our primary interest lies in the ordering of

[9] We are of course not considering problems of reliability in the sense of repeated observations of the same attributes. For convenience, we are tacitly assuming perfect reliability.

people, not the relative frequency of each position.

It might be asked how can one know the universe forms a scale if all one knows is a sample from the universe. At present it seems quite clear that in general the probability of finding a sample of attributes to form a scale by chance for a sample of individuals is quite negligible, even if there are as few as three dichotomous items in the sample and as many as one hundred individuals.[10] It seems quite safe to infer in general that if a sample of attributes is selected without knowledge of their empirical interrelationships and is found to form a scale for any sizeable random sample of individuals, then the universe from which the attributes are selected is scalable for the entire population of individuals.

### 13. SCALING AND PREDICTION

It is important to distinguish between two closely related topics, scaling and prediction. Finding that a universe of attributes is scalable for a population means that it is possible to derive a quantitative variable from the multivariate distribution such that each attribute is a simple function of that variable. We might phrase this otherwise by saying that each attribute is (perfectly) predictable from the quantitative variable.

This is the converse of the ordinary problem of prediction. In an ordinary problem of prediction, there is an outside variable, independently defined, that is to be predicted *from the attributes*. For example, it might be desired to predict the income of a student five years after he graduates from college, from his present knowledge of mathematics. To do this, an experimental sample would have to be obtained where salaries five years after college are known for each person and where responses to each item on the mathematics test are known. If the criterion of least-squares is adopted, then the best prediction on the basis of the sample would be the multiple regression of income on the three items in the sample. The multivariate distribution of the three items and the outside variable would give the necessary data for computing the regression, curvilinear or linear, which would be best for predicting the outside variable. If we wished to predict some other outside variable from the same items, a new multiple regression would have to be worked out from the multivariate distribution of the three items and the new outside variable. In general, the first of these regressions would ordinarily be expected to differ from the second. In general, weights to be used to predict one outside variable from a set of attributes will differ from those used to predict another outside variable; a new multiple regression must be worked out for each outside variable.

This emphasizes an important property of scales. If the items have a multivariate distribution that is scalable, it can easily be seen that no matter what the outside variable may be, the same prediction weights may be given to the items. The correlation of any outside variable with the scale scores is precisely the same as the multiple correlation of that outside variable with the items in the scale. Thus we have an outstanding property of scaling, namely, that *it provides an invariant quantification of the attributes for predicting any outside variable.* No matter what prediction purpose is to be served by the attributes, the scale scores will serve that purpose.

### 14. ON "ITEM ANALYSIS"

Let us repeat the distinction just made. In scaling we reproduce the attributes from a quantitative variable. In prediction, we predict a variable from the attributes. This is a sharp difference which enables us to avoid much of the confusion that seems to

[10] *Technical Footnote.* To work out the complete probability theory would require two things: first, a definition of a sampling process for selecting items, and second, a definition of what is meant by a scale not existing. A definition of the sampling process is difficult because items are ordinarily developed intuitively. Stating a null hypothesis that a scale does not exist leads to many possible analytical formulations, for different limiting conditions may be imposed upon the multivariate distribution of the items. For example, should the marginal frequencies be considered fixed in all samples, should the bivariate frequencies be considered fixed, etc.? These are questions which may become clearer as the theory of scaling develops, and in return may clarify our conceptions of what observation of social phenomena implies.

prevail in the previous literature on scale construction. It seems to have been felt that items in a universe are merely stepping stones from which to obtain scores. It seems to have been felt that it was an embarrassing deficiency to lack a particular variable to predict from the items—that as a necessary evil one had to resort to methods of internal consistency to derive scores.

This accounts for current "item analysis" approaches to scaling. These use procedures that are typically as follows. A trial set of weights is assigned the categories, yielding a trial set of scores. Then each item is examined to see how well it by itself discriminates between these scores, that is, how well the scores can be predicted *from the item.* Those items which individually discriminate best are retained, and the others eliminated.

The misleading character of such procedures can be seen by inspection of the examples of scales in §10 and §11 above. We have pointed out that the intercorrelations between attributes in a scale can be as close to zero as one pleases. It can also easily be seen that the correlation ratio of the scale *scores* with any single item can also be as close to zero as one pleases. The predictability of the scale variable from an attribute does not tell whether or not the attribute is predictable from the scale variable.

The use of the "item analysis" procedures in connection with scales seems to be an unfortunate carry-over from the problem of ordinary prediction of an outside variable. In such a prediction problem, the items are truly but stepping stones to enable predictions to be made. It is known[11] that item analysis affords a first approximation to multiple correlation (or the discriminant function), and an item is of interest only insofar as it aids in the multiple regression.

Our emphasis for scaling is quite different. In scaling, we are interested in each and every attribute in the universe on its own merits. If we were not, we would not work with the universe. The attributes are the important things; and if they are scalable, then the scores are merely a compact framework with which to represent them.

If a compact framework is found, it has the additional important property of being an efficient device for predicting any outside variable in the best manner possible from the given universe of attributes.

### 15. THE RELATIVITY OF SCALES

An interesting problem associated with scales is: why does a universe form a scale for a given population? For example, take the sample of three mathematics questions given above. Why should these three questions be scalable? There is no necessary logical reason why a person must know the area of a circle before he can know what a derivative is, and in particular the derivative of $e^x$. The reason for a scale emerging in this case seems largely cultural. Our educational system is such that the sequence with which we learn our mathematics in our high schools and colleges is first to get things such as areas of circles, then algebra, and then calculus. And the amount of drill that we have on each of these topics is probably also in that order. It would be quite possible, however, for the proverbial "man from Mars" to come to this earth and study calculus without having to learn the area of a circle, so that he might not be a scale type, according to the scale presented above; or a student may have had some personal incident which somehow impressed upon him with great force the derivative of $e^x$, but in the ordinary course of circumstances would have forgotten it even more readily than he forgot the area of a circle.

The scale analysis will pick out such deviants or non-scale types. Of course, if these non-scale types are too numerous, we shall not say that a scale exists. In practice we find scales, although never perfect scales, only because there has been sufficient uniformity of experience for the population of individuals so that the attributes mean essentially the same thing to the different individuals. As a matter of fact a study of the deviants is an interesting by-product of the scale analysis. Scale analysis actually picks out individuals for case studies.

A universe may form a scale for a popula-

[11] See, for example, Louis Guttman, "An Outline of the Statistical Theory of Prediction," in Paul Horst, et al., *The Prediction of Personal Adjustment,* Social Science Research Council, 1941.

tion at a given time and may not at a later time. For example, the items in the scale of expression of desire of American soldiers to go back to school after the war may not prove to be scalable if they were asked once more at the close of the war.

A universe may form a scale for one population of individuals, but not for another. Or the attributes may form scales for two populations in different manners. For example, a sample of items of satisfaction with Army life which formed a scale for combat outfits in the Air Force did not form a scale for men in the technical schools of the Air Force. The structure of camp life for these two groups was too different for the same items to have the same meaning in both situations.

If a universe is scalable for one population but not for another population, or forms a scale in a different manner, we cannot compare the two populations in degree and say that one is higher or lower on the average than another with respect to the universe. They differ in more than one dimension, or in kind rather than in degree. It is only if two groups or two individuals fall into the same scale that they can be ordered from higher to lower. A similar consideration holds for comparisons in time. An important contribution of the present theory of scaling is to bring out this emphasis quite sharply.

## 16. SUMMARY

1. The multivariate frequency distribution of a universe of attributes for a population of objects is a scale if it is possible to derive from the distribution a quantitative variable with which to characterize the objects such that each attribute is a simple function of that quantitative variable.

2. There is an unambiguous meaning to the order of scale scores. An object with a higher score than another object is characterized by higher, or at least equivalent, values on each attribute.

3. There is an unambiguous meaning to the order of attribute values. One category of an attribute is higher than another if it characterizes objects higher on the scale.

4. It can be shown that if the data are scalable, the orderings of objects and of categories are in general unique (except for direction). Both orderings emerge from analysis of the data, rather than from *a priori* considerations.

5. The predictability of any outside variable from the scale scores is the same as the predictability from the multivariate distribution with the attributes. The zero order correlation with the scale score is equivalent to the multiple correlation with the universe. Hence, *scale scores provide an invariant quantification of the attributes for predicting any outside variable whatsoever.*

6. Scales are relative to time and to populations.
   a. For a given population of objects, a universe may be scalable at one time but not at another, or it may be scalable at two periods of time but with different orderings of objects and categories.
   b. A universe may be scalable for one population but not for another, or it may be scalable for two populations but with different orderings of objects and categories.
   c. Comparisons with respect to degree can be made only if the same scaling obtains in both cases being compared.

7. From the multivariate distribution of a sample of attributes for a sample of objects, inferences can be drawn concerning the complete distribution of the universe for the population.
   a. The hypothesis that the complete distribution is scalable can be adequately tested with a sample distribution.
   b. The rank order among objects according to a sample scale is essentially that in the complete scale.
   c. The ordering of categories in a sample scale is essentially that in the complete scale.

8. Perfect scales are not found in practice.
   a. The degree of approximation to perfection is measured by a *coefficient of reproducibility*, which is the empirical relative frequency with which values of the attributes do correspond to intervals of a scale variable.
   b. In practice, 85 percent perfect scales or better have been used as efficient approximations to perfect scales.

9. In imperfect scales, scale analysis picks out deviants or non-scale types for case studies.