CHAPTER 1

Numerical data and the meaning of measurement

> There is no safety in numbers, or in anything else.
> (James Thurber)
>
> He that forsakes measure, measure forsakes him.
> (Fergusson's Scottish proverbs)

There is a myth about the way in which science works. Scientists attempt to find out how natural systems work. From what little they can glean initially by unaided observation and by analogy with what they think are similar systems, they form hypotheses about the workings of the systems they are studying and put these hypotheses to the test. They do this by making predictions from their hypotheses and then checking these predictions via observations aided by scientific methods, which may include experiments, measurements, and so on. These methods often involve elaborate equipment, stringent controls and highly standardised procedures. Because of their sophistication they are thought to provide a transparent window on reality. They show us how things really are. Observations made, the predictions can be checked against the data and science moves a step forward: the hypothesis is confirmed or falsified and this general procedure repeated. In this way, it is thought, science moves ever closer, by successive approximations, to an understanding of how natural systems work.

Like many myths, this one contains some truth. But if research in the history and philosophy of science over the past half-century has shown anything of value, it has shown that the methods that scientists use to test their hypotheses are not transparent windows on the world. Philosophers and historians divide over what kind of 'windows' these methods might be. Some think they are like the windows of Chartres cathedral, where what is seen is located within the window itself and not in the world beyond. Others think

that while there are distortions and discolouration, something of the world behind the window can, at least sometimes, be glimpsed. Obviously, the latter view is the most that any kind of research could ever force us to, for while methods may contaminate observations, possibly in ways we do not suspect or cannot easily see, blanket scepticism about methods of knowing is self-refuting and, incidentally, likewise defeating for the historian or philosopher of science as well.

The lesson to be learned is that scientific methods are imperfect tools and all observations are, in principle, fallible. Because scientific methods are imperfect, the only safe way to use them is critically. By this I mean that caution in science requires investigating one's methods as well as using them. I have heard scientists disclaim the need for this, arguing that one does not need to know how a car works in order to drive it. That might be true around the city, but try driving across Australia's Simpson Desert, without roads, on unchartered territory, without knowing how your car works! The scientist in the classroom giving demonstrations to students is like the driver in the city; the scientist in the laboratory, investigating as yet untested hypotheses, is like the driver in the desert.

The critical investigation of methods has two parts: empirical and conceptual. Any observational method, even 'naked' observation, because it involves a causal process between the observer and the observed, presumes a theory about how that method discloses some of nature's secrets to us. A good example is the way theories of optics underwrite the use of the telescope. These theories need to be empirically investigated, just like any others in science. But deeper than the empirical lies the conceptual underpinnings of methods. The critical investigation of methods, and their proper use, requires conceptualising the method correctly. If we consider an entire class of methods, such as methods of measurement, the conceptual problem resides in defining the method. This is neither a trivial nor an arbitrary exercise. Methods are interwoven inextricably into the fabric of science and the definition given of a concept such as measurement must be consistent with its place in that fabric. It is possible that uncritical scientists in a particular area could, for socio-historical reasons, come to misunderstand a concept such as measurement and use it in ways inconsistent with its wider theoretical commitments.

Then the methods called 'measurement' within that science would not disclose the sort of facts about the world that they might be thought to and those scientists would misunderstand what they were doing.

Modern psychology, quantitative and experimental, began with the publication in 1860 of *Elemente der Psychophysik* by the German scientist, G. T. Fechner. A physicist preoccupied with psychological questions, Fechner was guided by the uncompromisingly imperialistic metaphysical vision of natural science. In proposing a feasible scientific theory about how any natural system works, a metaphysical promissory note is thereby contracted, the scope of which encompasses all natural systems connected spatio-temporally, however distantly, with the system theorised about. This promissory note entails that the categorial features presumed in that theory infuse the spatio-temporal realm entirely. Categorial features are the warp and weft of being, so general that they permeate every situation, no matter where, no matter when. Two such, of fundamental importance to theories in physics, are causality and quantity. The category of causality underwrites the experimental method, that of quantity, measurement. These methods, experiment and measurement, are often seen as marks ratifying true science, and so are automatically imposed upon newer areas of scientific investigation. This was the case with Fechner's psychophysics, delivered already swaddled in measurement and experiment.

If quantity is present in every situation, it may seem that measurement is required of all sciences. Not so. This issue is more complex than at first appears, particularly in the case of psychology. The relationship between quantity, as a category of being, and measurement, as a method of science has never been rigorously examined. The founding fathers of modern psychology, almost to a man, simply presumed that measurement was a scientific imperative and, accordingly, thought to contrive quantification. Whether they were correct or not is a matter requiring careful analysis.

Can the existence of psychological measurement be seriously questioned, now, at the close of the twentieth century, with psychology so long and (seemingly) securely established as a quantitative science? Is it not a fact that psychologists measure an array of psychological attributes? Certainly, psychologists claim to be

able to measure such an array: psychological attributes like general intellectual ability ('intelligence'); various specific intellectual abilities (verbal ability, spatial ability and so on); the intensities of different kinds of sensations (loudness, brightness, etc.); the subjective probability of occurrence of various possible events (such as winning some gamble); the strength of attitudes towards social policies (e.g., euthanasia or abortion); the subjective value of various commodities (such as laptops or wilderness areas); degrees of personality traits (introversion, neuroticism, etc.); strength of association between a stimulus and the overt response elicited (such as Hull's 'habit strength'); levels of skill (e.g., social skill or typing skill); and levels of achievement in various areas (such as spelling or arithmetic). Not only psychologists, but the wider community accept that psychologists measure at least some of these. But science as knowledge, as distinct from science as a social movement, is often indifferent to the confidence of scientists and the vicissitudes of popular opinion.

In fact, there are signs that this presumption of successful psychological quantification is premature. One very disturbing sign is that many psychologists misunderstand what measurement is. In taking over the concept of measurement from the established sciences and fashioning their own quantitative theories and practices, psychologists are, like all scientists, logically committed to the traditional view of measurement; but in endorsing and promoting their claim to measure, psychologists typically invoke a definition of 'measurement' at odds with the traditional view.

The claim that psychologists measure psychological attributes is embedded in a complex matrix of concepts and practices. This matrix has three dimensions. First, there is an observational dimension: the sets of observational and analytical procedures applying, according to the relevant theories, to each such attribute. Second, there is a theoretical dimension: the character that each supposedly measurable attribute is taken to have, both its intrinsic character (i.e., how different levels of this attribute interrelate) and its extrinsic character (i.e., how the attribute relates to others). Third, there is a philosophical dimension: the understanding of measurement professed, in virtue of which psychologists think of their practices as measurement. There is a dissonance between these dimensions, a dissonance largely unac-

knowledged. It can be revealed via a brief examination of some examples of psychological 'measurement'.

For half a century or so after the publication of Fechner's *Elemente der Psychophysik*, psychophysics remained the principal area within which psychological quantification was attempted. During the twentieth century interest in psychophysics waned and attempts to measure intellectual abilities became the central focus of quantitative psychology. The technology of ability measurement, so-called, is perhaps the most significant contribution, for better or worse, that modern psychology has made to our society. The examples considered in this chapter, accordingly, are taken from these two areas.

First the observational dimension will be examined, then the theoretical. What kind of thing is it that psychologists suppose they are able to measure? In discussing this question, interest will not be in how sensation intensity and intellectual ability should, separately, be defined. Instead, it will be in the general character they are thought to share in virtue of being hypothesised as measurable. Within psychology, it is supposed that sensation intensity and intellectual ability are both quantitatively related to other attributes. Theorising of this sort carries implications about the internal character of the attributes involved, and these, in turn, entail a view of measurement.

The definitions of measurement which psychologists typically present in their publications will then be considered. It transpires that the definition of measurement entailed by the theory and practice of psychology is quite different from the definitions which psychologists explicitly profess. It will be argued that in formulating their own, special definition of measurement, psychologists undermine the understanding of measurement implicit in the theories they propose and on which their quantitative practices depend.

TWO EXAMPLES OF PSYCHOLOGICAL MEASUREMENT

Much of what passes for psychological measurement is based upon the counting of frequencies. A sequence of situations is constructed, each of which delivers just one of two possible outcomes via the behaviour of participants, some of whose attributes it is

intended to measure. Responses to mental test items, classified either as correct or incorrect, is a typical example. The number of outcomes of one kind or the other is counted, and a mathematical theory linking these frequencies (or, perhaps, associated probabilities) to the attributes to be measured is accepted as true and, via that theory, measures of those attributes obtained. This pattern is easily replicated for many different psychological attributes because dichotomous situations are easily contrived. The following examples display this pattern.

Psychophysics

The aim of psychophysical measurement, as conceived by its founder, Fechner, is to quantify the intensity of sensations. Consider a set of stimuli, all of which vary only with respect to a single, directly discernible, quantitative, physical attribute (e.g., a set of spherical marbles, all of the same colour and volume, but varying in weight). The idea behind Fechner's psychophysics was that the presentation of each stimulus gives rise to a mental state, a sensation (e.g., the sensation of heaviness produced by a marble held in the palm of the hand) and it was thought by Fechner, and many after him, that the physical magnitude of the stimulus and the intensity of the corresponding sensation are related by some mathematical formula (or function), one specific to that attribute (e.g., weight)[1]. There are a variety of procedures by which it is thought that the intensity of sensations can be measured. The instance considered here is the method of pair comparisons.

This method involves presenting elements from the stimulus set, two at a time, to the person whose sensations are being measured, with instructions to report for each pair which of the two is the 'greater' in some prescribed sense (e.g., which of two marbles is the heavier). This procedure is repeated many times under standard conditions, including repetitions of the same pair. Ideally the procedure is continued until a relatively stable estimate is

[1] Of course, Fechner recognised that the same stimulus presented to the same person on different occasions but under otherwise identical external conditions could give rise to sensations of different intensity, in which case what may be said to correspond psychologically to the physical magnitude of the stimulus is a probability distribution over a range of possible sensation intensities. The mathematical psychophysical functions which Fechner and others proposed were intended to relate to something akin to the average of these distributions.

obtained of the person's probability of judging, under these conditions[2], each particular stimulus, say x, greater than each other stimulus, y. Alternatively, if it is thought that individual differences in the sensations produced are not too great, then repeated observations on the same stimulus pairs may be obtained from different people[3]. Over many repetitions of each stimulus pair, the number of times x is judged greater than y in the required sense may be counted for all pairs, x and y, and these frequencies converted to proportions. These proportions or, more correctly, the probabilities that they are thought to reflect, are taken to vary systematically with the magnitude of the difference between the sensations produced by the stimuli involved. If the precise relationship between such probabilities and differences were known as a general law, then the sensation differences could be measured via the proportions. Such a law cannot be known *a priori*, but a number of mathematical relationships have been hypothesised. Perhaps the best known is L. L. Thurstone's Law of Comparative Judgment (Thurstone, 1927a, b).

In the 1920s and later, Thurstone's theoretical work in psychological measurement synthesised and organised many of the previously disparate ideas in the area. One important idea which had not been explicitly developed theoretically until then was the idea that magnitudes of the relevant stimulus attribute do not unvaryingly cause fixed intensities of sensation.[4] A stimulus of a given magnitude (e.g., a marble of a particular weight), says Thurstone, may give rise to any of a range of sensations, some being more likely than others. Here Thurstone employs the so-called Normal (or Gaussian) probability distribution form:[5] the probability

[2] It is now recognised that the relevant conditions are not only physical but psychological. In particular, motivational and cognitive factors are known to be important.

[3] Repetitions of pair comparisons involving the same stimulus pairs can be very boring for subjects, and so can have a dramatic effect upon motivational states. This is just one of the difficulties in making such measurements which I shall ignore in developing this example.

[4] This will be because of small fluctuations in the state of the causally relevant parts of the nervous system which cannot be experimentally controlled with existing technology and not, of course, because of any intrinsic indeterminism in the sensory system.

[5] This distribution form is named after the German mathematician, Karl Friedrich Gauss (1777–1855). Its form is that of the now familiar bell-shaped frequency distribution, widely referred to in many sciences where statistics are analysed, especially the biological, behavioural and social sciences. It was already very well known in psychology when Thurstone proposed his theory and in that context, some adjudge it a not implausible hypothesis (see Luce, 1977, 1994).

distribution of sensation intensities associated with stimulus x is Normal with mean, μ_x, and variance, σ_x^2. Thus, the expectation is that x will produce a sensation in the vicinity of μ_x, the likelihood of something too much greater or less than μ_x diminishing with distance from that mean value, the extent of diminution being dependent upon the magnitude of σ_x^2. Similar expectations hold for each other stimulus, say y, where the relevant parameters are μ_y and σ_y^2 respectively. Simplifying by assuming no response biases (i.e., that the person making the judgment uses a simple response rule[6]), that for all stimulus pairs, x and y, $\sigma_x^2 = \sigma_y^2$ and that the sensation intensities elicited by x and y on any occasion are independent of one another, Thurstone's 'law' becomes

$$z_{xy} = \frac{\delta_{xy}}{\sigma} \qquad (1)$$

(where z_{xy} is the Normal deviate[7] corresponding to $P_{x>y}$ (the probability of judging x greater than y), $\delta_{xy} = \mu_x - \mu_y$, and σ is the standard deviation (i.e., the square root of the variance) of the distribution of differences in intensity between sensations elicited by any one stimulus in the set and those by any other, a constant for all stimulus pairs under the simplifications assumed here. For convenience, the unit of measurement can be set at σ, in which case σ equals 1). This mathematical relationship can be described approximately in less mathematical terms as follows: when $\delta_{xy} = 0$ (i.e., when $\mu_x = \mu_y$), $P_{x>y} = .5$; as δ_{xy} increases from 0, $P_{x>y}$ increases from .5, at first rapidly approaching 1, but never reaching it because the rate of approach gradually and continually slows down; and as δ_{xy} decreases from 0, a mirror-image process happens, with the probability now approaching, but never reaching, 0.

Taking the proportion of times that x is judged greater than y as an estimate of $P_{x>y}$, this probability can be transformed to z_{xy} and the difference between μ_x and μ_y, δ_{xy}, can, accordingly, be estimated. When repeated for all pairs in the stimulus set, measures

[6] Thurstone assumed this simple response rule: if on any occasion the intensity of the sensation produced by x exceeds that produced by y then the person involved will always judge x greater than y. Subsequently, it has been thought that people may not always behave in this straightforward way.

[7] The Normal deviate corresponding to a probability is the point under the standard Normal curve (i.e., the Normal curve with a mean of 0 and a variance of 1) below which the proportion of the total area under the curve equals that probability.

of the μ values for all stimuli can be estimated on a scale with arbitrary zero point (see Bock & Jones (1968) for an account of the estimation procedures). If Thurstone's conjecture and the simplifying assumptions are true, then the estimated μ values can be interpreted as the expected value of the sensation intensity produced by the stimulus involved under the kind of observational conditions employed. Here, then, is one case of putative psychological measurement.

Intellectual abilities

It is a commonplace observation that people differ in their performances on intellectual tasks. Two people invited to solve an arithmetic reasoning problem, for example, will often give different solutions. This fact has been used, at least since the work of Binet (1903) and Spearman (1904) to attempt to measure intellectual abilities. Intellectual abilities are hypothesised properties of persons which are supposed to be responsible for differences in performance on intellectual tasks. Of course, such differences in performance will have a variety of causes within the persons involved, not all of them intellectual. For example, it is widely believed that motivational factors play a part. Intellectual abilities are usually thought of as distinct from such factors, having to do exclusively with what the person involved knows (the person's cognitive state) or with the neural mechanisms sustaining such knowledge. Since the time of Spearman a variety of theories has been proposed for the measurement of intellectual abilities.

These theories typically apply to scores on psychological tests. Such tests are fixed sets of intellectual problems administered under relatively standardised conditions. The individual problems involved are called test items. When administered to a person, the person's solutions to the items (that person's responses) are recorded. These responses are then classified as *correct* or *incorrect* and, typically, the number of correct responses (the person's total score) is the datum from which a subsequent measure of ability is inferred. That is, it is generally thought that intellectual abilities relate in a systematic way to total scores. It is pertinent that the relationship between the sets of item responses and total scores is not one to one. Obviously, with the exception of the two possible extreme scores on any test, two people could get the same total

score by getting different items correct. Thus, if total scores relate systematically to intellectual abilities, the relationship between abilities and item responses must be less systematic, because it must be possible for two people, having exactly the same level of ability and getting the same total score,[8] to get different items correct. Most theories in the area cope with this requirement by postulating a probabilistic relationship between abilities and item responses.[9] Current theories of this sort are called *item response theories*.

Item response theories connect the probability of a person getting a test item correct to some combination of the person's ability and attributes of the item. The relevant item attributes are usually taken to be the difficulty of the item, the discriminating power of the item and the probability of getting the item correct by random guessing. To illustrate the measurement of intellectual abilities using this approach, imagine a test in which differences in total scores between people depend only upon differences between them in one ability. Of course, this is an idealisation, but it may be approximated in the case of certain simple tests.

An item's difficulty level is located on the same scale as the person's ability: the difficulty for any item is the level of ability required to have a 50:50 chance of getting the item correct. If a person has less ability than this, the chances of failing the item should exceed those of passing it, and if they have more ability they are more likely than not to pass it. This last feature relates to the item's discriminating power. The more rapidly the probability of getting an item correct increases as ability increases above the item's difficulty (or the more rapidly this probability falls away with decreases in ability below the item's difficulty) the better the item discriminates between different levels of the relevant ability.

Suppose now that the probability of getting an item correct on this imaginary test varies with just two attributes: the person's level of the relevant ability, and the item's difficulty (all items having the same discriminating power). Each item classifies

[8] Of course, two people with the same level of ability need not get the same total score either.

[9] Again, the probabilistic relationship need not be taken as implying indeterminism. It no doubt simply reflects the failure of the psychometrician to control all relevant causal factors.