

3

Levels of Measurement

Chris Fife-Schaw

- 3.1 Introduction
- 3.2 Classifying measurements
 - 3.2.1 *Categorical measures*
 - 3.2.2 *Ordinal level measures*
 - 3.2.3 *Interval level measures*
 - 3.2.4 *Ratio scale measures*
- 3.3 Discrete versus continuous variables
- 3.4 Measurement errors
- 3.5 Choices over levels of measurement
- 3.6 The relationship between level of measurement and statistics
- 3.7 Conclusions
- 3.8 Further reading

AIMS

This chapter introduces the reader to the common categories of measurement used in psychological science. This traditional categorization system is fundamental to understanding how to conduct good research but is also key to making decisions about how to analyse the data generated by a study. The chapter also briefly describes some challenges to this orthodox view.

Key terms

approximate value	interval measurement
categorical measurement	mutual exclusivity
conjoint measurement	non-parametric tests
theory	ordinal measurement
continuous variables	parametric tests
discrete variables	ratio scale measurement
exhaustiveness	real limits

3.1 INTRODUCTION

While there are many aspects of the research process that do not involve measurement and, indeed, some fields of research where explicit measurement is avoided altogether, the great majority of research studies will involve it in some form. Whether a research hypothesis stands or falls may depend on how well the key concepts have been measured, independently of whether or not it is a worthy hypothesis. What follows in this chapter is a discussion of measurement issues that have been central to the pursuit of 'positivist' psychological science. Key amongst the assumptions (see Cattell, 1981) is that before we can construct grand psychological theories and laws, we must first be able to measure and describe things with reasonable accuracy.

For the purposes of this chapter, measurement is defined as the assigning of numbers to objects, events or observations according to some set of rules. Sometimes these numbers will be used merely to indicate that an observation belongs to a certain category; at other times these numbers will indicate that the observation has more of some property than an observation that is assigned a lower number.

In much of psychology we have to measure psychological properties indirectly because we have no direct access to the mental constructs we want to measure. It is a straightforward matter to measure length and we can do this fairly directly by offering up our measuring instrument (ruler or tape measure) to the object we want to measure. In the case of IQ, however, we can only infer levels of intelligence from tests that ask people to solve problems of varying difficulty. We assume that people who get more of the more difficult items correct are more intelligent, but we cannot yet observe intelligence in any more direct way than this. In many respects the existence of something called intelligence is itself a hypothesis and the debate about what IQ tests *actually* measure has often been a heated one in the past. While few people would argue about what a ruler measures, the quantities measured by many psychological measurement instruments are more open to debate and much more obviously depend on the theoretical perspective of the researcher than is the case in the physical sciences.

This is not to say that psychological measurements are of little value. A great deal of effort has been expended in establishing the reliability and validity of psychological measures over the last century. There are now libraries of well-validated tests for all sorts of psychological phenomena which can be used very effectively as long as the manuals are used appropriately. Chapter 10 outlines the principles involved in test construction and development commonly used in psychology.

While many established tests exist, researchers are often confronted by the need to create their own measures to deal with the specific problems they have. This may be because nobody has yet fully developed a test for the particular

kinds of observations you are interested in. It may be that you are measuring something which has not been measured before or, perhaps, that the existing tests are too cumbersome for your purposes. Here, you will have to pay particular attention to the precise meaning and nature of your new measures.

It should go without saying that the goal is always to measure things as well as possible. There are often trade-offs that have to be made, however. Measures that demand lots of time and effort from participants may induce fatigue and boredom that may simply introduce unwanted 'noise' into your measurements. On the other hand, measures that are very simple and quick for the respondents to complete are frequently crude and inaccurate. Ultimately you will have to make the judgement as to whether your measures are 'good enough' for your purposes.

3.2 CLASSIFYING MEASUREMENTS

Whether you use a ready-made measure or create your own, you always need to know what class of measurement you have made. How you classify a measurement will have an impact on the kinds of numerical analyses you can perform on the data later on. Stevens (1946) proposed that all measurements can be classified as being of one of four types. This system has become dominant within psychology and no methods textbook would be complete without describing it. There are, however, other important alternative conceptualizations of measurement such as those of Luce, Krantz, Suppas and Tversky (1990) and Adams (1966), and objections to the way psychologists think (or rather, do not think) about their measures (see Box 3.1). Stevens's classification remains the best known but it is only one way of thinking about measurement.

3.2.1 Categorical measures

Categorical measurements (variables), also called nominal measurements, reflect qualitative differences rather than quantitative ones. Common examples include categories such as yes/no, pass/fail, male/female or Conservative/Liberal/Labour. When setting up a **categorical measurement** system the only requirements are those of mutual exclusivity and exhaustiveness. **Mutual exclusivity** means that each observation (person, case, score) cannot fall into more than one category; one cannot, for example, both pass and fail a test at the same time. **Exhaustiveness** simply means that your category system should have enough categories for all the observations. For biological sex there should be no observations (in this case people) who are neither male nor female.

A key feature of categorical measurements is that there is no *necessary* sense in which one category has more or less of a particular quality: they are simply

categorical measurement

Mutual exclusivity

Exhaustiveness

different. Males are different from females (at least at some biological level) and northerners come from the north and southerners do not. Sometimes, however, this will seem like an odd assumption. Surely 'passing', for example, is better than 'failing'? Well, yes, in certain cases this would be so, but this would depend on what your a priori theory about the measure was. If you believed that 'passing' was more valuable and reflected more positively on somebody (e.g. that they were more intelligent, or paid more attention) then that is a matter for you as a researcher; the use of a pass/fail category system does not inherently contain any notion of greater or lesser value.

For the purposes of using computers to help with our analyses, we commonly assign numbers to observations in each category. For instance we might assign (code) a value of 1 for males and 2 for females. The important point is that although females have a numerically larger number there is no suggestion that being female is somehow better or more worthy. Again, this can cause confusion, especially as your computer deals only with numbers and not their meanings. You could, for instance, ask it to calculate the mean sex of the respondents and it might come up with a figure like 1.54; clearly this is pretty uninformative other than that it tells you that there are more females than males.

Although the categories of a categorical variable do not necessarily have any value associated with them, this does not mean that they cannot reflect some underlying dimension in some instances. As an example, you might classify people you are observing in the street as 'young' or 'old' because you are unable to approach them to ask their ages directly. While this is likely to be an extremely crude and inaccurate classification, this system implies an underlying continuous dimension of age even though we place people in only two categories.

The criteria for categorical measurement do not preclude the possibility of having a category of 'uncategorizable'. If you were to have such a category you would satisfy both the mutual exclusivity and exhaustiveness criteria, but if there were a lot of 'uncategorizable' observations then the value of your categorization system might be brought into question. How useful is it to have a variable on which the majority of observations are 'uncategorizable'? This can only truly be answered with reference to your research question.

3.2.2 Ordinal level measures

ordinal measurement **Ordinal measurement** is the next level of measurement in terms of complexity. As before, the assumptions of mutual exclusivity and exhaustiveness apply and cases are still assigned to categories. The big difference is that now the categories themselves can be rank-ordered with reference to some external criterion such that being in one category can be regarded as having more or less of some underlying quality than being in another category. A lecturer might be asked to rank-order their students in terms of general ability at statistics. They could put each

student into one of five categories: excellent, good, average, poor, appallingly bad. Clare might fall into the 'excellent' category and Jane into the 'good' category. Clare is better at statistics than Jane, but what we do not know is just how much better Clare is than Jane. The rankings reflect more or less of something but not *how much* more or less.

Most psychological test scores should strictly be regarded as ordinal measures. For instance, one of the subscales of the well-known Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975) is designed to measure extroversion. As this measure, and many like it, infer levels of extroversion from responses to items about behavioural propensities, it does not measure extroversion in any direct sense. Years of validation studies have shown how high scorers will tend to behave in a more extroverted manner in the future, but all the test can do is rank-order people in terms of extroversion. If two people differ by three points on the scale we cannot say *how much* more extroverted the higher-scoring person is, just that they are more extroverted. Here the scale intervals do not map directly on to some psychological reality in the same way that the length of a stick can be measured in centimetres using a ruler. The fundamental unit of measurement is not known.

Since many mental constructs within psychology cannot be observed directly, most measures tend to be ordinal. Attitudes, intentions, opinions, personality characteristics, psychological well-being, depression, etc. are all constructs which are thought to vary in degree between individuals but tend only to allow indirect ordinal measurements.

This conclusion is a point of contention for many researchers since one of the implications of assuming these measures to be ordinal is that some parametric statistical tests should not be used with them. Indeed, even the humble mean is not used appropriately with ordinal measures (the median is a more appropriate measure of central tendency). This sits uneasily with what you will see when you read academic journal articles, where you will regularly find means and parametric statistics used with ordinal measures. We will deal with this issue later in this chapter (see also Chapter 19).

3.2.3 Interval level measures

Like an ordinal scale, the numbers associated with **interval measurement** reflect more or less of some underlying dimension. The key distinction is that with interval level measures, numerically equal distances on the scale reflect equal differences in the underlying dimension. For example, the 2°C difference in temperature between 38°C and 40°C is the same as the 2°C difference between 5°C and 7°C.

interval measurement

As we will see later, many behavioural researchers are prepared to assume that scores on psychological tests can be treated as interval level measures so that

they can carry out more sophisticated analyses on their data. A well-known example of this practice is the use of IQ test scores. In order to treat scores as interval level measures, the assumption is made that the 5-point difference in IQ between someone who scores 75 and someone who gets 80 means the same difference in intelligence as the difference between someone who score 155 and someone who scores 160.

3.2.4 Ratio scale measures

Ratio scale measurement

Ratio scale measurement differs from interval measurement only in that it implies the existence of a potential absolute zero value. Good examples of ratio scales are length, time and number of correct answers on a test. It is possible to have zero (no) length, for something to take no time, or for someone to get no answers correct on a test. An important corollary of having an absolute zero is that, for example, someone who gets four questions right has got twice as many questions right as someone who got only two right. The ratio of scores to one another now carries some sensible meaning which was not the case for the interval scale.

The difference between interval and ratio scales is best explained with an example. Suppose we measure reaction times to dangers in a driving simulator. This could be measured in seconds and would be a ratio scale measurement, as 0 seconds is a possible (if a little unlikely) score and someone who takes 2 seconds is taking twice as long to react as someone who takes 1 second.

If, on average, people take 800 milliseconds (0.8 seconds) to react we could just look at the *difference* between the observed reaction time and this average level of performance. In this case the level of measurement is only on an interval scale. Our first person scores 1200 ms (i.e. takes 2 seconds, 1200 ms longer than the average of 800 ms) and the second person scores 200 ms (i.e. takes 1 second, 200 ms more than the average). However, the first person did not take 6 times as long (1200 ms divided by 200 ms) as the second. They did take 1000 ms longer, so the *interval* remains meaningful but the ratio element does not.

True psychological ratio scale measures are quite rare, though there is often confusion about this when it comes to taking scores from scales made up of individual problem items in tests. We might, for instance, measure the number of simple arithmetic problems that people can get right. We test people on 50 items and simply count the number correct. The number correct is a ratio scale measure since four right is twice as many as two right, and it is possible to get none right at all (absolute zero). As long as we consider our measure to be *only* an indication of the number correct there is no problem and we can treat them as ratio scale measures.

If, however, we were to treat the scores as reflecting ability at arithmetic then the measure would become an ordinal one. A score of zero might not reflect absolutely no ability at all as the problems may have been sufficiently difficult so

that only those with a moderate degree of ability would be able to get any correct. It would also be a mistake to assume that all the items were equally difficult. Twenty of the questions might be easy and these might be answered correctly by most people. Getting one of these correct and adding one point to your score would be fairly easy. The remaining items may be much more difficult and earning another point by getting one of these correct might require much more ability. In other words, the assumption that equal intervals between scores reflect equal differences in ability is not met and we should strictly treat the scores as an ordinal measure of ability. Even when doing this we are assuming that ability is a quantitative entity though we will not have established this directly (see Box 3.1).

Box 3.1 Are we deluding ourselves about our measures? A word of caution

Recent years have seen a challenge to the orthodoxy on measurement presented in this book, most notably by Joel Michell (e.g. Michell, 2000). Michell's arguments are highly detailed philosophical ones and it is difficult to represent them fairly in a short space; however, a key idea in his work is that in the rush to appear to be 'hard' scientists like physicists, psychologists, and psychometricians in particular have failed to consider some fundamental questions about what they are assuming when they attempt to measure psychological attributes. When coming up with a quantitative measure of some attribute psychologists are assuming that the attribute concerned has a quantitative structure, yet this is rarely, if ever, tested – even though, Michell argues, that this is in principle an empirical question open to investigation. If trying to measure job satisfaction, say, psychometricians rarely stop and ask the question 'is job satisfaction really a quantitative attribute?' – it is already assumed to be quantitative and indeed it is necessary to assume this if the quantitative test scores are to have any sensible meaning. The focus usually moves directly to how satisfaction test scores are quantitatively related to other variables, even though the quantitative nature of satisfaction was never established. Satisfaction could be a categorical state for instance, and it is far from proven that dissatisfaction is the dimensional opposite of satisfaction.

The existence of a test that produces numbers does not establish that the attribute being 'measured' is really quantitative and a lot of bogus 'science' may be built on flawed measurement assumptions. Although Michell speculates about why psychologists and psychometricians have not bothered with establishing that given attributes are quantitative, doing so is not a simple matter. **Conjoint measurement theory** (e.g. Luce & Tukey, 1964) offers one of the few ways to address this at the moment, and Michell (2000) gives a nice illustrative example. Other methods have proven elusive, yet the need for them is clear – we should not be attempting to present psychology as a rigorous science that measures quantitative things if we cannot establish that the things we want to measure are actually quantitative in the first place.

**Conjoint
measurement theory**

As hopefully it will have become clear, there is a hierarchical distinction between the types of measurement described in this section. Nominal measures give information on whether two objects are the same or different, ordinal measures add information concerning more or less of a quantity, interval measures add information on the distance between objects, and ratio scale measures add the absolute zero standard.

3.3 DISCRETE VERSUS CONTINUOUS VARIABLES

discrete variables

Many types of measurement result in indices that consist of indivisible categories. If someone scores 13 on our 50-item arithmetic test, they might have scored 14 on a better day but they could never have scored $13\frac{1}{2}$. The score $13\frac{1}{2}$ was not possible as the individual questions can be marked only correct or incorrect. Measures like this are called **discrete variables** since they can have only discrete, whole number values.

continuous variables

Some variables such as height and time are referred to as **continuous variables** since they could be divided into ever smaller units of measure. We could measure height in metres, then centimetres, then millimetres, then micrometres, then nanometres and so on until we got to the point where our measuring instrument could not make any finer discrimination. There are an infinite number of possible values that fall between any two observed values. Continuous variables can be divided up into an infinite number of fractional parts. Ultimately it is the accuracy of our measuring instrument that puts limits on the measurement of continuous variables. If our ruler can measure accurately only to the nearest millimetre we must settle for that degree of precision.

approximate value

When measuring a continuous variable you end up recording a single figure, but this really represents an interval on the measurement scale rather than a single value. It is therefore always an **approximate value**. If we time someone doing a task to the nearest second, and it takes them 20 seconds, we are really saying that the time taken lies somewhere in the interval between 19.5 s and 20.5 s. Had it actually taken them 19.4 s we would have rounded the time to 19 s, not 20 s (note: to avoid rounding bias when rounding a number that ends exactly with a numeral 5, round to the nearest even number). Similarly, an elapsed time of 20.6 s would have been rounded to 21 s.

In this example we are deliberately recording times only to the nearest second but, in principle, the choice of any measurement tool carries with it a limit to the degree of accuracy that can be achieved and thus the rounding process will have to happen even if we are unaware of it. We will still be reporting a time that corresponds to an interval and not a discrete value. If our stopwatch could record times to the nearest hundredth of a second, say, and we recorded a time of 20.12 s, this would still mean we were saying that the time taken lay somewhere in the

interval between 20.115 and 20.125 s. These boundary values are referred to as **real limits**.

real limits

It is always appropriate to use the most accurate measure practicable. Any calculations you do using approximate values necessarily include that approximation in the final result. Use two or more approximate values in a calculation and the scope for misleading results increases dramatically. Hence it is always preferable to use approximate measures associated with the smallest intervals possible so as to minimize this problem. You should also note that, although our variables might be theoretically continuous, such as time and length, the act of measurement always reduces the measure to a discrete one.

3.4 MEASUREMENT ERRORS

The goal of all researchers should be to minimize measurement errors. Put formally, these are the discrepancies between the observed value of your measurement and the 'true' value. There is a simple formula to illustrate this:

$$\text{Observed score} = \text{True score} + \text{Error.}$$

The 'error' term may be positive or negative. Obviously it would be nice to have the error term as small as possible. If you were measuring people's heights with a ruler marked off in inches then you could probably only measure accurately to within half an inch. Having a ruler marked off in millimetres would give rise to much more accurate measurement, and finer distinctions between individuals could be made (see the previous section). In a similar way, psychological measures should strive to make as fine a set of distinctions between people as possible. Assuming your measure is valid, it makes sense to have more points on your measurement scale rather than fewer.

This holds true only so long as you believe the individual points on the scale carry the same meaning for all participants. When it comes to ratio scale and interval level measures, this is not a problem. You could measure time to the nearest millionth of a second, though you might find the necessary timing equipment a little expensive! For most psychological research, timing to the nearest millisecond is probably accurate enough. Things get much more difficult when you have ordinal measures, however. Problems arise when you try to label individual responses on your ordinal scale. Take the following as an example.

Let us assume you have an attitude statement about a political issue and you would like people to tell you how much they agree or disagree with it. You could provide a five-point scale as follows:

1	2	3	4	5
Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree

Most respondents would know what they were required to do with such a response scale. While you could not be certain that all those who 'agreed' had agreed to the same extent, you would probably feel reasonably happy that they did not intend to tell you they had very strong views on the topic. Similarly, it is probably safe to assume that they are not entirely equivocal about the issue either.

If you gave this question to several hundred people in a survey, however, you might find that so many people had the same score on the item that it did not discriminate very much between people. In this situation you might want to increase the number of response options available. A seven-point scale could be used and it would be reasonably easy to label the response options. You might even think a nine-point scale was appropriate, though labelling all the points might prove more of a challenge. Indeed, you could simply label the end- and mid-points, leaving the rest unlabelled.

Why not opt for a 29-point scale instead? This would give even greater discrimination, surely? The answer is, regrettably, no. Respondents would now have trouble working out where they should indicate their response on the scale. Should it be the 18th or the 19th point or even the 20th? Such a response format increases the scope for confusion on the part of the respondent and thus will introduce, rather than reduce, measurement error. There is also the problem that we still do not know that all people responding at point 19 agree to the same extent. Such multi-point ordinal scales introduce an unfortunate illusion of precision.

3.5 CHOICES OVER LEVELS OF MEASUREMENT

In the previous and very traditional section you will have noticed that I have implicitly suggested that ratio and interval level measurement is to be preferred over ordinal or categorical measures. The reason for this is that in most cases a good ratio scale measure will contain more information about the thing being measured than a good ordinal measure. You would probably rather have temperature reported in degrees Celsius than on a scale of very cold, cool, neither warm nor cold, warm, very hot. You should always strive for greater accuracy of measurement where possible.

Naturally some kinds of variable are always going to be categorical (e.g. sex) and some are always going to be ordinal (e.g. most scaled measures; but see Section 3.6). In such cases you should not regard your measures as somehow inferior. Whilst it would be nice to think that ultimately we will have access to

more direct measures of attitudes and personalities, for example, these are not likely for the foreseeable future.

There are, however, some common practices which should be discouraged. The most notorious of these is the collapsing of ordinal measures into categorical ones. It is quite common to see researchers take an attitude item with a seven-point agree/disagree response format and collapse the data into a simple three-point scale of agree/uncertain/disagree. This practice degrades the measurement by removing the extremity information.

There are three kinds of motive for collapsing data in this way. One is the desire to use simpler statistical procedures; a second is to make graphs and tables clearer; and the third is that you might not believe that your seven-point measure is very accurate or valid. With the ready availability of comprehensive statistics books and computer programs the first problem is easily overcome. While clarifying graphs and tables is an admirable aim, it would be desirable to collapse the scores only for this purpose and conduct statistical analyses on the uncollapsed data. The third justification is also a justification for not using the measure. If you doubt the validity or accuracy of a measure then you should think twice about using it at all.

3.6 THE RELATIONSHIP BETWEEN LEVEL OF MEASUREMENT AND STATISTICS

Most good statistics texts present 'decision trees' which help you select the correct statistical test to use providing you know the answers to a number of simple questions about your data and research design. These are very useful, and simple versions are provided in Chapter 19 on bivariate analyses.

These decision trees ask about the level of measurement for your data as well as the nature of the distribution of scores on the measure that you expect in the population from which your sample scores were drawn. The topic of distributions of scores is dealt with in Chapter 19 but the level of measurement issue is pertinent here, particularly at the boundary between ordinal and interval level measures.

The attraction of **parametric tests**, ones that assume something about the distribution of scores in the population (e.g. *t*-test, ANOVA), is that there are many more of them than **non-parametric tests**. They often allow you to ask interesting questions about your data that are not easily answered without using such parametric procedures. To say that your measure is only ordinal, rather than interval level, usually rules out these useful procedures. Chapter 20 outlines some of the many possibilities. Two views have developed over the appropriateness of treating ordinal measures as interval ones. Those interested in reading more on this debate should see Henkel (1975), Labovitz (1975), Davison and

parametric tests

non-parametric tests

Sharma (1990), Townsend and Ashby (1984) and Stine (1989), among many others.

One view states that, most of the time, providing you have a good-quality ordinal measure, you will arrive at the same conclusions as you would have using more appropriate tests. It is sometimes argued (see Minium, King & Bear 1993) that while most psychological measures are technically ordinal measures, some of the better measures lie in a region somewhere between ordinal and interval level measurement.

Take a simple example of a seven-point response scale for an attitude item. At one level this allows you to rank-order people relative to their agreement with the statement. It is also likely that a two-point difference in scores for two individuals reflects more of a difference than if they had only differed by one point. The possibility that you might be able to rank-order the magnitude of *differences*, while not implying interval level measurement, suggests that the measure contains more than merely information on how to rank-order respondents. The argument then runs that it would be rash to throw away this additional useful information and unnecessarily limit the possibility of revealing greater theoretical insights via more elaborate statistical procedures.

The more traditional and strict view (e.g. Henkel, 1975; Stine, 1989) says that using sophisticated techniques designed for one level of measurement on data of a less sophisticated level simply results in nonsense. Computer outputs will provide you with sensible-looking figures but these will still be nonsense and should not be used to draw inferences about anything. This line of argument also rejects the claim that using parametric tests with ordinal data will lead to the same conclusion *most of the time* on the grounds that you will not know when you have stumbled across an exception to this 'rule'.

The debate on this issue continues. The safest solution, advocated by Blalock (1988), is to conduct analyses on ordinal measures using both parametric and non-parametric techniques where possible. Where both procedures lead you to the same substantive conclusion then, when reporting parametric test results, you will at least know that you are not misleading anyone. You should be guided more by the non-parametric procedures if the conclusions are contradictory.

What would be unacceptable is to select the statistical procedure that leads to results that support your hypothesis. You should attempt consistency in reporting findings so that you decide either that your data meet the assumptions for parametric procedures or that they do not.

Ultimately, whether this issue matters will depend on the seriousness of making a mistake and who your audience is likely to be. Research on a drug or an intervention that may change people's lives demands the most strict and conservative approach to your analysis. On the other hand, if your research topic is more esoteric and your audience is researchers in a field that has regularly used (abused?)

parametric techniques on ordinal data, then you may find it difficult to get a hearing if you do not report findings in the accepted way.

3.7 CONCLUSIONS

This chapter has attempted to alert you to the main issues surrounding levels of measurement. As time marches on, the research community may come to an alternative system of classifications (cf. the debate discussed above). However, the Stevens system described here remains the dominant one in psychology for the time being. Chapter 19 takes this a step further by looking at the principles of statistical inference in more detail. Be sure that you have understood this chapter before you read Chapter 19.

3.8 FURTHER READING

All good statistics textbooks explain Stevens's classification system and the relationship levels of measurement and statistics, though few books will go much beyond what has been presented here and in Chapter 19. Minium *et al.* (1993) has the virtue of spelling out many of the debates in a clear and accessible way. Many of the key papers on the debate about measurement and statistics have appeared in the *Psychological Bulletin* and are likely to continue to appear in that journal.