CHAPTER 2: MEASUREMENT IN PSYCHOLOGICAL RESEARCH

LEARNING OBJECTIVES:

- Understand conceptual issues in measurement for psychological sciences.
- Understand the distinction between levels of measurement for observed data.
- Build an awareness of how decisions on measurement can influence study conclusions.

INTRODUCTION

How tall are you? How wide is that table? What is the temperature today? These are all questions about a measurement, and all seem relatively straightforward to answer. How extraverted is John? This question is also about a measurement, but we hope it is immediately clear that this is less straightforward to answer.

In this chapter, we will introduce the conceptual challenge of measurement in the psychological sciences, and attempt to illuminate exactly what that final question is so much more difficult for us to answer. Next, we will introduce the classic definitions of levels of measurement originally proposed by Stevens (1946), and discuss the properties of data that conform to these definitions. A particular emphasis throughout this chapter will be placed on the way choices about measurement can influence the conclusions that we may draw from data. Finally, we will introduce some further terminology that is commonly used in describing data, and that will carry into the practical language used when working with data in R.

BROAD PRINCIPLES OF MEASUREMENT

Before we begin a caveat. Measurement in psychological science is a much debated topic. What constitutes measurement, what can and cannot be measured, the philosophical status of what is measured, the nature of measurement scales, are all current topics. In the discussion that follows, we will present a short summary, which will barely scratch the surface of such debates, but will generally set the scene for this book (see chapter reading list for introductory resources to these debates).

As was discussed in the previous chapter, our research questions should, if formulated well, provide us with some key constructs of interest, and a statement about the relationship between them. By constructs, we are referring to the psychological phenomena that we wish to study. Often in psychological science, these constructs are not directly measurable. Consider for example the concept of well-being. There is no real single "thing" which we might call well-being, and as a result, we need to define it more precisely, before we can attempt to measure it. We might choose to define well-being as a feeling of psychological happiness, or of physical health, or a combination of both. These definitions give rise to ways in which we might think about measuring well-being in a given sample, and thus collecting data on it.

So for example, if we defined well-being as psychological happiness, we might ask people to rate their own levels on happiness on a scale, or we might ask others to rate those close to them (or both). These ratings would produce us some data points on a variable we might call well-being. Alternatively, if we chose to define well-being as a state of physical health, we might measure an individual's blood pressure, resting heart rate, maximal oxygen uptake (VO2 max), cholesterol etc., and then find some way to combine them into a single measure of physical health. The "scores" or data points for well-being in both cases would represent very different things, and it would be difficult to call either *the* way of measuring well-being, yet both would seem reasonable to some degree.

It is useful to when thinking about defining the construct we wish to measure and the associated data, to draw an explicit distinction between the construct, the measure, and the variable we ultimately analyse. Figure 2.1 displays a proposed relationship between these three related ideas.

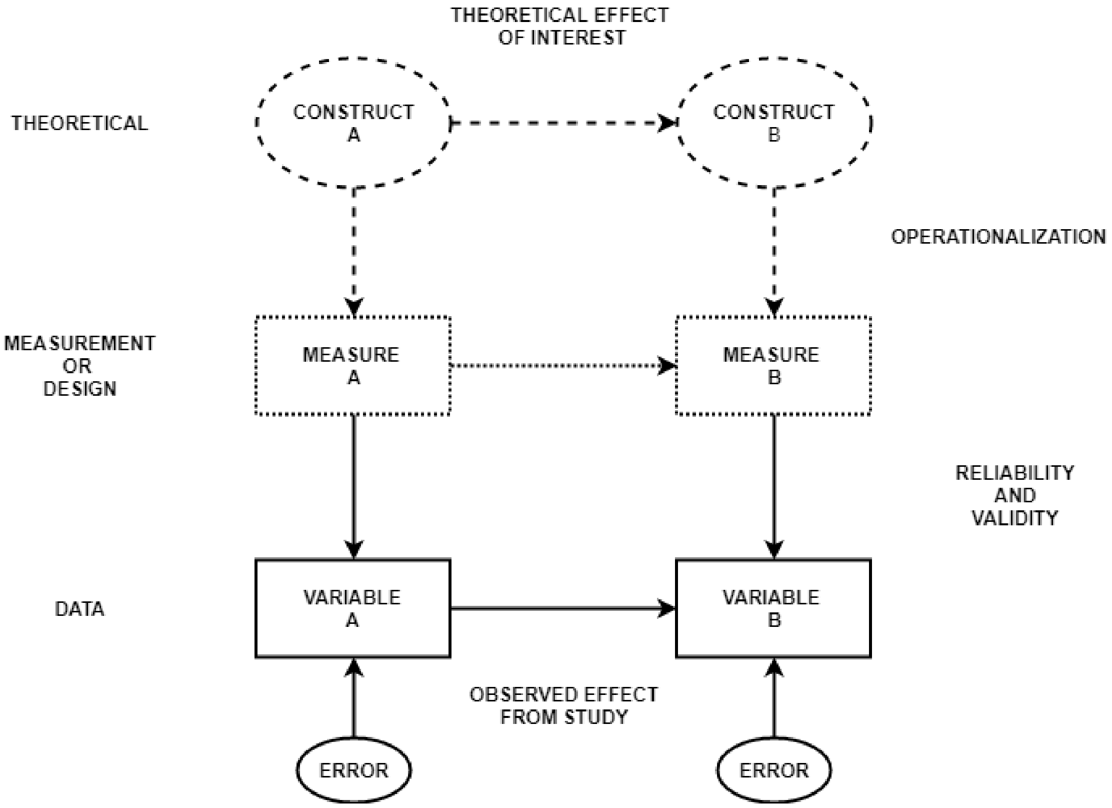

FIGURE 2.1: GRAPHICAL REPRESENTATION OF THE RELATION BETWEEN CONSTRUCTS, MEASUREMENTS AND VARIABLES.

When we pose research questions, we are usually doing so with reference to constructs we deem to be important. These constructs are depicted as dashed ellipses in Figure 4.1. Our research question, if well formed, should include a number of constructs (here A and B), and a relation between them. Here the relation between them is a single headed arrow, used to indicate that construct A predicts construct B. Our questions and hypotheses relate to this relationship. However, we do not directly observe this relationship. The effect we observe in our data is a number of steps removed.

First, we need to choose how we will measure our constructs. These measurements come from our decisions about the particular definition of the construct we wish to use, and then the associated choices of study design. The measures may be a particular survey, experimental paradigm etc. The process of selecting a specific measurement tool for our constructs of interest is referred to as the **operationalization**. There exists a relationship between the construct and its operationalization, depicted by the vertical dashed arrows pointing downwards from the ellipses in Figure 4.1. In selecting a particular measure, we are assuming that it closely represent the construct or process we are studying. However, it is important to be aware that this is an assumption, an assumption that often cannot be tested. It is the choice of the researcher.

Ideally, the operationalization (measure) will relate as closely as possible to the theoretical definition of the construct. The tightness of this relationship varies across sub-disciplines in psychology. For example, the field of psychometrics, which plays a critical role in individual difference psychology among others, concerns the measurement of things like personality, well-being, intelligence etc. many of which are theoretically abstract. However, other areas of psychology make use of phenomena such as reaction time, which is arguably less ambiguous to measure – although its exact meaning can still be a topic of debate.

A common practical approach in applied research is to use measures that have been widely used in previous studies. However, whilst this may reduce the variation in researcher choices of measurement instrument, it does not speak in any way to the reasonableness of the link between the construct and the measure. Indeed, if a researcher is working from a different definition of the theoretical construct than the original developers of a measure, the use of the same measurement instrument may be inappropriate.

Once we have selected a method of measurement, we can think about the relationship between measure A and measure B, depicted by the dotted arrow in the middle of section of Figure 2.1. We can think about this as an approximation of our relationship of interest that is *in principle* possible given our choice of measurements. We use in principle here because, when we conduct a study, we apply our measures to samples of participants. This results in measured variables. The individual values for each participant on those measured variables is our study data. These variables are depicted as the solid rectangles at the bottom of Figure 2.1.

As with our previous step, there is a relationship between the measure and the resultant variable. However, this relationship we are able to assess to some degree by evaluating the reliability and validity of our measures. In brief, **reliability** concerns the accuracy in our measurement. **Validity** concerns the degree to which our measurement tool measures what it was designed to measure. We will discuss reliability and validity in more depth in Chapter XX, but they are important concepts in measurement. In general, we will always want our measures to be as accurate as possible. It is not very useful to us if we have a highly reliable measure of something we have no interest in, or equally if we have a very unreliable measure of something we are interested in – both are suboptimal.

Now, the final step in our journey is to consider the solid arrow drawn between variable A and variable B at the bottom of Figure 2.1. This is the *observed effect*. By observed effect, we mean the value of some statistic we have calculated to represent the theoretical effect (dashed arrow), in other words, the result of our study. The important point here is that we base the conclusions we draw in a given study on this observed effect. What we are able to say about the theoretical effect is dependent on the quality and plausibility of our steps in operationalizing and measuring our constructs.

Our ability to say *anything at all* is also going to be dependent on the amount of error that exists in both our overall process, and our specific measurements. We are going to see throughout this book that in conducting statistical analyses, we are usually making a comparison between an observed effect and error, of which measurement error is an important part. Measurement error is depicted by the solid ellipses with arrows pointing to variables A and B. These arrows are showing that error in our measurement directly influences the quality of the data we have on our variables. If the measurement error is too high, when compared to the size of the effect we are trying to measure, then we are unlikely to get reasonable results. The methodologist Andrew Gellman (http://andrewgelman.com/2015/04/21/feather-bathroom-scale-kangaroo/) used the following analogy (slightly adjusted here) to make point.

Suppose we are trying to measure the weight of a feather. To do so, we use a set of bathroom scales, and place the feather in the pouch of a kangaroo that is jumping up and down. In this analogy, the weight of the feather is the observed effect. The bathroom scales and kangaroo would be our (very bad) design, which introduces a huge amount of error. The effect we are interested in – the weight of the feather – is tiny compared to the measurement noise/error – the bathroom scales and the jumping kangaroo. In general, we want to have as little error in our study as possible. Therefore, at the point of study design we need to pay careful attention to whether our choice of measurement is precise as possible.

LEVELS OF MEASUREMENT AND DATA TYPES

How we choose to measure our constructs has another very practical implication, it produces different types of data. We describe, visualize and analyse data of different types in different ways. So it is important that we think about the type of data the will result from a given design, and the subsequent properties this data might have, in order that we can appropriately plan our analyses.

As a brief example, suppose we are studying performance at university. We want to gather data on students marks. The university we have selected to gather data from keep student records in two formats, as the raw scores (number of correct answers) on all assessments, and as grade – i.e. whether the student is getting an A, B, C etc. Now, say I look at data from three participants on one assessment. Table 2.1 contains the data.

TABLE 2.1: TEST PERFORMANCE FOR THREE PARTICIPANTS

Participant	**Score**	**Grade**
Participant 1	35	A
Participant 2	29	A
Participant 3	28	B

Now, suppose we ask you the following question: *Using either the Score or Grade data in Table 2.1, which two students are most similar in their level of performance on the test?* Take a minute to think about the answer to that question based on each set of data.

Hopefully it is clear that the answer to the question depends on which of the columns we look at. If we consider the Score, the number of correct answers, we would likely conclude that participants 2 and 3 are most similar as they are only one point different in their scores, and all other differences are larger. However, if we were to look at Grade, we might conclude that participant 1 and 2 are most similar, as they have the same grade, and both differ from participant 3. In this example, because we can see both sets of data side by side, the difference in the conclusions we would draw are obvious. However, if we had chosen to only measure performance based on Grade, then we would not have had this comparative information available to us, and it would be far less clear whether any such interpretational differences exist.

The preceding example informally touches on a number of points we will return to in this book about the importance of our decisions on measurement, and how this relates to the information we have at our disposal for answering our research questions. In short, we will always want to have the most information rich variant of a measure possible. What do we mean by information rich? Consider the two examples above again. In the case of Score, we know exact marks. These marks will in principle, though not necessarily in any given class, range from the minimum possible, to the maximum possible. Therefore, the Score variable contains pretty much all the information we would want on the test. In contrast, the Grade variable groups students into categories. Within a particular grade boundary, we do not have any information about student test results. That is, all we know is that participants 1 and 2 received an A, and we will know the range of scores that constitute an A. However, that is all we know, and it would not be possible for us to gain any more information from that variable.

The example also points to some fundamental differences with respect to nature and type of data our design and measurement choices yield. That is, when discussing correct answers, we could think concretely about the differences in the number of correct answers. In thinking about Grade, we think more about whether individuals are grouped together or not. What this leads us towards is the idea that data have different properties that influence the types of ways we can think about, and analyse it. The properties associated with data define their **levels of measurement.**

LEVELS OF MEASUREMENT

In Stevens (1946) taxonomy, there are four types of data; nominal, ordinal, interval and ratio. As we move up the levels of measurement, the numeric values associated with an observation convey increasing amounts of information.

Nominal data differentiates participants based on the membership of groups. So for example, consider the top panel in figure 2.2. Say we ask people what city they live in. We would get a variety of answers (e.g. London, Edinburgh etc.) and we could assign each of these cities a number. However, the numbers assigned to the cities do not have any mathematical meaning. Like the names of the cities themselves, all they say is that anyone with a specific number (referred to here as a level) of the variable (City) belongs to the same group. As a result, it does not matter which number is assigned to which city, it is arbitrary. In the example in figure 2.2, Dublin could just has easily been given the numeric value 1 as 4. All we are able to ask based on the values assigned to nominal variables is whether the value for one participant is the same as the value for another participant, or to put it another way, whether participants belong to the same group.

The most common use of nominal variables in psychology is to assign people to experimental groups (see Chapter 3, Study Design), and when we collect demographic information on participants.

Ordinal data conveys meaning about relative position. The numeric values assigned to ordinal data, as the name suggests, allows for the responses to be placed in a specific order. So, if we consider the example in figure 2.2, suppose we ask a question about how much someone likes chocolate, and we provide a response scale with the verbal labels shown below the scale. We assign each of these verbal labels a number from 1 to 5 which represents the degree of positive liking for chocolate.

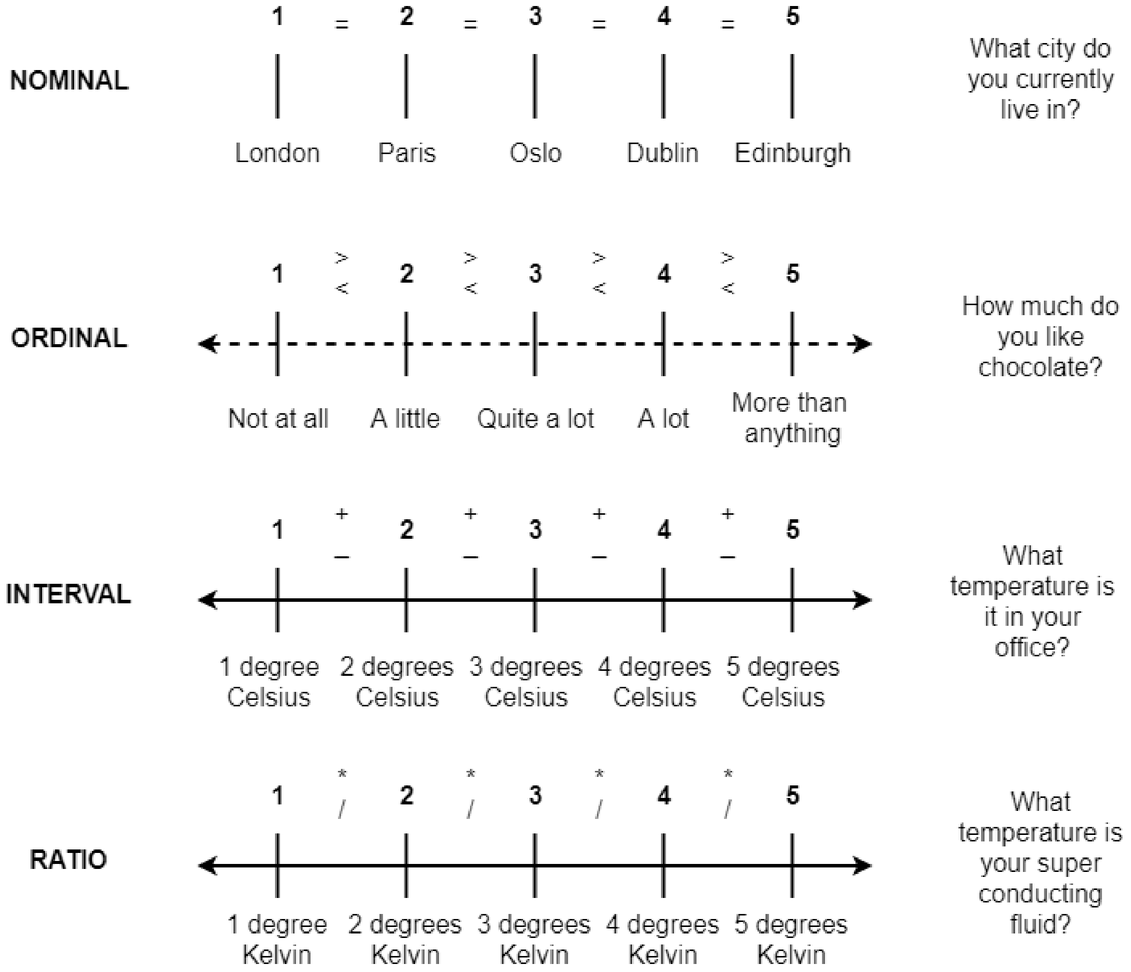

FIGURE 2.2: REPRESENTATION OF LEVELS OF MEASUREMENT, ASSOCIATED MATHEMATICAL OPERATIONS AND EXAMPLE QUESTIONS.

From such a question, we may well be happy to say that Mary who responds with a 2, likes chocolate more than Pete who responds 1, and that Dave who responds 3, likes chocolate more than Rachel who responds 2. Both Mary and Dave are one point higher on the response scale than Pete and Rachel respectively. However, the differences between the numbers are not interpretable. That is, we could not say that Mary likes chocolate more than Pete does, to the same extent that Dave likes chocolate more than Rachel does, even though the numerical difference between the numbers is one in each case.

This is a little easier to think about if we replace the numbers with the verbal labels. If we were to ask, is the difference between “Not at all” and “A little”, the same as the difference between “A little” and “Quite a lot”, you would probably all intuitively think *what a stupid question*, or more perhaps more politely, *how on earth do I answer that? Maybe!* In short, adding numbers to labels does not change the nature of the measure we have chosen, and doesn’t by definition make numerical questions reasonable.

Ordinal level data is very common in many areas of psychology. For example, a huge proportion of measures of constructs such as personality, use rating response formats similar to those in Figure 2.2 – often referred to as Likert-type scales (often incorrectly as Likert scales – but this is a very specific type of scale). As we will discuss later in this book, how such scales are analysed often assumes they have properties above ordinal status – but we will come back to this.

When data measured at the **interval** level, not only can we discern the ordering of observations, but also the differences between the observations are also meaningful. The classic example of interval data is temperature – as shown in the third panel of figure 2.2. Temperature in degrees Celsius (and also Fahrenheit) are interval measurements. The difference between 1 degree C and 2 degrees C, is the same as the difference between 3 degrees C and 4 degrees C. It is therefore meaningful for us to say that there is an equal *difference* in temperature between 1 and 2 and 3 and 4 degrees – it is a difference of 1 degree. This is very useful to us when it comes to analysing data. Think back to our discussion of the types of question that we might ask in conducting a psychological study, some of these concerned differences.

However, with interval level measurement, it is not possible for us to make statements like *"4 degrees Celsius is twice as hot as 2 degrees Celsius".* This is because interval data does not have a meaningful zero point. If we think about temperature, what does zero degrees C mean? Well, degrees Celsius are scaled such that this is the temperature at which water freezes. But this does not mean 0 temperature, or a point at which there is an absence of temperature. We can contrast this with length. Zero centimetres means no length. So the value zero for length in centimetres and temperature in degrees Celsius are fundamentally different.

Length in centimetres is an example of measurement at the **ratio level.** Ratio data has all of the properties of the previous levels of measurement, but it has a meaningful zero point. Although degrees Celsius is not a ratio variable, degrees Kelvin (see Figure 2.2) is. Zero degrees Kelvin (-273.15 degrees Celsius) is known as absolute zero, is considered to be the lowest possible temperature. Importantly, when scales have a true zero, we are able to meaningfully discuss ratios of variables. Therefore, while it is not meaningful to state that 4 degrees Celsius is double 2 degrees Celsius, it is meaningful to say that 4 degrees Kelvin is double 2 degrees Kelvin.

Practically it is important to remember that the properties of data we are describing here, are the properties related to the operationalization of a construct using specific measures as part of our design. As a simple example, consider measuring age. We could operationalize age by asking people to respond to the question "*What is your age*?" by ticking which of a set of options (e.g. 18-30; 31-40; 41-50, etc.) applies to them. Alternatively, we could operationalize age by asking for date of birth. The former operationalization would result in a variable at the ordinal level, the latter operationalization a variable at the interval scale. In short, levels of measurement are not inherent in *constructs*, but are properties of measures. ***Our choices matter.***

SOME CAUTIONS

There are many published critiques of Stevens classifications, and these began to appear very soon after the original publication. The critiques vary, from an initial concern about the lack of a formal mathematical basis, to the incompleteness of the taxonomy. The incompleteness rears its head repeatedly when we try to think through examples of variables and how they may fit into the taxonomy. Here the Stevens classifications are useful as they get us thinking in more depth about the relations between our study decisions and our data, and that is really our primary goal in this chapter. It is important to remember that very often in data analysis there is not cut and dry, right and wrong answer. This can be one of the hardest things to get used to when you begin your data analytic learning.

If you would like to spend some more time thinking about levels of measurement, look at the section of the on-line material "Examples of Measurement", where we discuss further typical measures in psychology and their classification.

YET MORE DEFINITIONS

There are two further sets of definitions, or classifications, of types of data that are useful to set out early, namely, the difference between **numeric** and **categorical** data, and **continuous** and **discrete data**.

Numeric, or numerical data, is data for which the numeric values assigned from a process of measurement carry some meaning. Thus, we tend to refer to interval and ratio level data as numeric. Categorical data, as the name would suggests, uses numeric values to assign observations to categories. Thus, we tend to refer to nominal and ordinal data as categorical.

Binary or **dichotomous** data is a special case of categorical data where there are only two categories.

Continuous data is data that can take on any value within a range. As a result, we can choose any two points on the scale of a continuously measured variable, and all real number values between those points would be possible to observe. An example will help clarify. Suppose want to measure the length of a table. We have a tape with centimetres marked on it, and the table measures in at between 45 and 46 centimetres. This would not worry us, because we know that "underneath" centimetres is a finer level of measurement, millimetres. Further, we know that millimetres are a meaningful metric of length, and that 45.2 centimetres, or 452 millimetres, is shorter than 45.8 centimetres, or 458 millimetres. Below the millimetre would be the micrometer, and so on. The point here is that we can split between any points, and still retain a numeric value which carries meaning. Practically, the main limitation is usually the precisions of our measurement tools.

In contrast, **discrete data** can only take on specific values and retain meaning. Another way to think about this is that we can not sub-divide the space between two points to come up with another number which is meaningful. The easiest way to think about discrete data is to think about counting the number of times an event occurs. Think about a count of the number of people who have crossed a road. Would it be possible for us to make a statement like, *"4.2 people crossed the road*"? This would seem very odd. So discrete data is countable, and can only take whole integer values.

Ratio and interval data can be continuous. In Figure 2.2, this this depicted by the solid lines running horizontally in the bottom two panels. Nominal and ordinal data are discrete. In

Figure 2.2, this this depicted by the absent or broken lines running horizontally in the top two panels.

SUMMARY

The current chapter introduced a number of key conceptual issues with measurement in psychological science, and highlighted where we make assumptions in the measurement process – some we can test, some we cannot. We discussed the classification of measurements into different levels and data types, and considered what these classifications mean with respect to the meaningful calculations on the resultant data.

WHERE NEXT…

THIS BOOK

Now we have introduced the fundamentals of data and measurement, we can move on to consider which types of data typically arise from different types of study.

IN LEARNING

As noted at the outset of this chapter, we have touched on a number of topics for which there is a lot of statistical, theoretical and philosophical debate. In the limited number of pages of an introductory data analysis textbook, there is simply not scope to do all of these debates justice. The reading list for this chapter is a little longer than for other chapters, and for the interested reader, should provide a number of accessible introductions to these extended debates.