Chapter 1
The Idea of Measurement

Measurement has a long history and is so familiar to us from our work in elementary measurement of length, mass, temperature and the like, that we take some of its intrinsic properties for granted. It is so familiar that it is expected that primary school children understand its key features. We consider these features in studying attempts at measurement in the social sciences. In the process, we need to distinguish between measurement and another familiar concept, that of *assessment*. The way we present it, assessment is a necessary precursor to measurement. The assessment provides the evidence which, under certain very strict conditions, may be transformed into measurements.

Latent Traits

Objects, persons, institutions or entities in general have properties which can be thought of in terms such as more or less, larger or smaller, stronger or weaker and so on. For example, people may be more or less able at English literature or mathematics, more or less neurotic, more or less for capital punishment, more or less tall and so on. Corporations may be more or less community friendly, schools may be more or less successful in producing leaders in society, and roads may be more or less accident prone. It is such *properties* of entities that are to be measured, not the entities themselves. A researcher does not measure a person but a psychological attribute (property) of the person such as neuroticism or intelligence.

In the social sciences, these *properties* are often also referred to as *constructs, attributes* and *traits*. All terms have similar connotations. In measurement and statistical contexts, where some kind of scoring is associated with the trait, the same idea is also referred to as a *variable*.

Thus, the following terms are more or less synonyms, with each having a slightly different nuance relevant for somewhat different contexts:

property ≅ trait ≅ construct ≅ attribute ≅ variable

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_1

Traits are also sometimes referred to as constructs because they are hypothetical and thus ‘constructed’ in order to be used in theories to explain human behaviour. You will pick up the different nuances in the use of these terms in context. However, be alert to the contexts so that you do become familiar with them as soon as possible.

Assessment: A Distinction Between Latent and Manifest

A trait is generally not measured directly. It is measured indirectly through its *manifestation*. Therefore, before a trait can be measured it is necessary to have a controlled procedure to manifest the property. This procedure we refer to as *assessment*. *Assessment* is a common term and we are using it essentially as it is commonly used in education and the social sciences. For example, it might be said that the mathematics proficiency of a person is *assessed* using a mathematics test, or that the neuroticism of a person is *assessed* using a neuroticism questionnaire.

In order to stress that traits are assessed indirectly through their manifestations, traits are often referred to as *latent*.

Thus, an assessment is a set of observations that arise when the manifestations of some property are observed in some systematic way that is acceptable to the research and professional field of expertise. These observations are often said to be produced from an *instrument* and in assessment of proficiency, where tasks to be solved are presented, they are generally referred to as *tests*. In the case of attitudes or opinions, where questions or statements are presented, they are often referred to as *questionnaires*. We already used these terms in the two examples above.

Scoring Assessments

The observations from an assessment, often referred to as *responses*, are qualitative. However, as a step towards measurement, these responses have an order which immediately implies more or less of the property to be assessed. In general, and immediately, this ordering is reflected by numbers assigned to the responses. In the case of tests of proficiency using the multiple choice items in which one response is deemed correct and all others incorrect, the incorrect and correct responses are scored 0 and 1, respectively. Clearly, the score of 1 reflects more proficiency than the score of 0. In the case of the neuroticism questionnaire, the four responses from strongly agree, agree, disagree through to strongly disagree may be scored 0, 1, 2 and 3, respectively. If the statement implies neurotic behaviour, then a strongly disagree response with a score of 3 will imply less neuroticism than a response of agree with a score of 1. This is ordering of a response and the assignment of an integer to characterize order is a significant, perhaps even a profound, step towards quantification, but it is not measurement.

We refer to responses assigned integers beginning with 0, putatively ordered responses, as *scored* responses. The step of scoring responses, which we generally take for granted, is the most important step. It is also perhaps surprising that Rasch's advanced mathematical theory leads to exactly this kind of intuitive assignment of successive integers to qualitative, but putatively ordered, responses to assessments. Many of the analyses that are carried out with responses, directly or indirectly, check whether this step has been carried out adequately. It pertains to both the reliability and validity of the assessments.

Dichotomous Items and Their Scoring

Instruments are generally composed of one of two kinds of items, or a combination. One kind is assessed simply *correct* or *incorrect*, and the incorrect response is scored 0 and the correct response scored 1. These are called *dichotomous* items and are said to be scored *dichotomously*. Clearly, there is a *direction* to the scoring. The response of 1 implies a better achievement or proficiency on the trait than a score of 0. In attitude assessment, the dichotomous responses might be the choice between *agree* and *disagree*. A decision has to be made as to which of the two responses is to be scored 1 and which is to be scored 0.

Polytomous Items and Their Scoring

The second kind of item permits assessment in more than two levels of proficiency. The scoring of these items is an extension of the scoring of the dichotomous items. Thus, the incorrect response is assigned a score of 0, and then successive levels of proficiency, or partial credit, are given successive integers until the maximum score is given to the totally correct response. They are called *polytomous* items and are said to be scored *polytomously*. An item assessing attitude might have the four responses, *strongly agree, agree, disagree and strongly disagree*.

We will see that different approaches to using these scored responses have different degrees of rigour in their approximation to measurement, with Rasch measurement theory (RMT) being the most advanced. The central part of this book is to learn how assessments may be carried out, how they may be transformed into measurements using RMT, and in the process, how to better understand the traits that are assessed and measured.

Key Features of Measurement in the Natural Sciences

Key features of measurement are that the trait can be mapped onto a line, often termed a *linear continuum*, and that the line can be divided into equal *units*, which can be made greater or smaller, from some origin. These units reflect the precision of the measuring instrument, with smaller units reflecting greater precision. This is clear with the measurement of length itself.

To measure a property of an object, the object needs to be engaged with or brought in contact with the measuring instrument. This engagement manifests the property of the object to be measured. This is a process for measurement, and using manifestations of properties of objects to measure them is not confined to the social sciences. Measurement in the natural sciences also requires the property of an object to be assessed to be manifested in some way.

Consider, for example, using the beam balance as a prototype for measuring the mass of an object. Objects with equal mass, that is units of mass, can be accumulated on one side until the beam balances the mass of the object on the other side. To represent this process, a line representing the continuum of mass can be drawn, and the mass of an object can be located on this line. Smaller units of mass give greater precision of the measurement. This is an example where the assessment instrument is so advanced scientifically that it can also immediately provide measurements. This advanced state of assessment instruments is a feature of the natural sciences. For example, spring balances, more complicated than the beam balance, and more recently electronic instruments, immediately give a reading of the mass in terms of units. The same feature is familiar in the measurement of temperature.

The presence of a direct reading of measurement from an assessment instrument in the natural sciences, such as a reading of mass or temperature, results in measurement and assessment often being taken together. Thus, the expression to measure something implies both the assessment and the measurement. When we need to construct or evaluate an instrument, as is often the case in the social sciences, we need to keep the distinction between assessment and measurement. Nevertheless, because of their close connection in the natural sciences, *to measure* is used even in the case where only the assessment step has been carried out.

In order for measurements to be meaningful, the instruments, their units and the origin have to be agreed to by those who use them. The history of physical measurement shows that the standardization of units in modern measurement is relatively recent (Alder, 2002). Understanding the traits in question and the factors that affect them is central to constructing measuring instruments. Attempts to construct measuring instruments, in turn, therefore, can clarify an understanding of a trait.

Stevens' Levels of *Measurement*

Because of attempts to clarify the meaning of measurement and how it might be applied in the social sciences, Stevens (1951) defined measurement as the assigning of numbers according to a rule. He also introduced the terms *nominal*, *ordinal*, *interval* and *ratio* levels of measurement. Already we can see a disagreement between Stevens' perspective and what we have said above. We have stated that the step of assigning numbers to assessments according to a carefully constructed rule in which a greater integer score reflects a greater value of the property, provides only the step of *scoring*, not measurement.

Unfortunately, despite its intentions to clarify the idea of measurement, many researchers in social science measurement consider that Stevens' classification system added to the confusion for the social sciences about the meaning of measurement, rather than a clarification. We mention his definition at the outset because you will come across it in readings in social science measurement. We now briefly review his four levels, but rather than referring to them as levels of measurement, we refer to them simply as a hierarchy in the use of numbers.

Nominal Use of Numbers

According to Stevens' definition, *nominal measurement* refers to assigning numbers only to indicate that, because two numbers are *different* from each other, then two objects assigned different numbers are also different from each other. Numbers on players' clothing in sports are generally of this kind. Because they are of this kind, carrying out standard numerical operations on these numbers does not produce numbers that are meaningful in the context. Therefore, it seems strange to refer to such assignment of numbers as measurement in any sense. We would say it is a nominal use of numbers, not nominal measurement.

Ordinal Use of Numbers

The numbers in *ordinal measurement* give only the *order* of the objects with respect to the trait. No inferences can be made regarding the size of the differences between objects. Our examples above of scoring an assessment are of this kind. Ranks also are of this kind. It is possible for the difference between successive ranks, which numerically is just a difference of 1, to represent much more variable differences on the trait. For example, a person ranked first may be very close to a person ranked second on some trait, or a great deal better than the person ranked second. Again we would say it is an ordinal use of numbers, not measurement.

Interval Use of Numbers

The numbers in interval *measurement* have a unit but an arbitrary origin. In this case, the differences between numbers on the scale are meaningful. For example, suppose an object is of mass 200 kg, but for some reason, the scale starts with a 0 at 100 kg. If the differences represent real differences of the properties of objects, for example, suppose on the new scale with the arbitrary origin at 100 kg, object A has the number 300 and object B has the number 200. Then the difference between the numbers for A and B is 100. However, object A is really of mass 400 kg and object B is really of mass 300 kg. The difference between their masses is indeed 100 kg.

However, ratios of the numbers are not meaningful. Thus we cannot infer that one object's size of its property is twice that of another object's just because the ratio of the respective numbers is 2. Take again the example of an object which is of mass 200 kg, but for some reason, the scale starts with a 0 at 100 kg. Then the number associated with the object is now 100. If we double this number, we get 200, suggesting that an object with twice the mass of the object is 200 kg. However, if we double the number of the actual mass of the object (which is 200 kg) we get 400 kg. Thus, doubling the number 100 does not give us the correct size of an object which is twice the mass of the original object.

The familiar Celsius or centigrade scale and, in some countries, the familiar Fahrenheit scale, for the measurement of temperature, are of this kind. Their origin, the 0 number on the scale, is arbitrary and does not represent a real temperature of 0. Through experimentation and theoretical developments, a real origin of 0 temperature is estimated to be −276.16 °C and −459.7 °F.

Researchers do refer to the application of numbers in the way described here as *interval level of measurement*. The reason it is reasonable to apply the term *measurement* here is that differences are meaningful in terms of a unit, and arithmetic operations, including ratios, can be carried out meaningfully on these differences.

Ratio Use of Numbers

The numbers, when assigned to properties of objects, in which *ratios* are immediately meaningful have both a natural origin and a defined unit. We can say, for example, that if the number assigned to object A is twice as large as the number assigned to object B, then object A has twice as much of the property as object B. For example, if object A is of mass 10 kg and object B of mass 20 kg, we can say that object B is 20/10 = 2 times the mass of object A.

We show in this book that with well-executed assessments with relevant scoring that have measurement in mind, we can approach measurement at the interval use of numbers. In principle, only the origin is arbitrary.

Reliability and Validity

The process of mapping the amount of a trait on a line which can give measurements necessarily involves numbers. The use of numbers in this way gives the potential for precision that is not possible with qualitative descriptions. However, just because they appear to be so precise, the precision can readily be over-interpreted. The topic concerned with degrees of precision and related issues is generally referred to as *reliability*. In addition, without a strong theoretical underpinning of the trait that is to be measured, the instrument may provide assessments, and hence apparent measurements, of a trait that is somewhat different from the one intended. This topic concerned with ensuring that assessments and measurements are of the trait intended is referred to as *validity*. Ideally, assessments and measurements are both reliable and valid.

You will already know that most tests and questionnaires are composed of many items. The reason for having many items, rather than just one, is to increase the precision and the validity of an assessment and measurement. Precision is potentially increased because there are more score points to distinguish the objects of assessment. Validity is potentially increased because each item can assess a slightly different aspect of the trait to be measured. When all items assess a common trait, and each assesses only its own unique aspect of the trait, then the assessment is said to be *unidimensional*. If different items of a test assess different traits and some different combinations of items assess different aspects of a trait, then unidimensionality of assessment is violated, and it may be said that the assessment is multidimensional.

We are concerned with constructing measurements that are unidimensional. However, we need to study our assessments to check if they are indeed unidimensional. They may be multidimensional and if they are, it becomes difficult to transform the assessments into a measurement on a single continuum. Of course, as you will see, unidimensionality is a matter of degree.

The term *construct*, which is one of our synonyms with trait above, emphasizes that the measurement of a trait is constructed. In doing so, it helps reinforce that this construction requires substantial experience and understanding. We revisit this term in Chap. 3.

As another preliminary note, we need to recognize that some important educational and social issues may *not* be readily amenable to measurement. One of the important functions of this book is to make you more able to construct and interpret measurements in education, health and the social sciences without falling into the many possible misunderstandings when using numbers as measurements.

Some Definitions

For the purposes of this book, *assessment* involves the engagement of an entity with some instrument, and the recording of observations of the engagement according to some protocol. *Measurement* involves some kind transformation of assessments and is defined as *the estimation of the amount of a unidimensional trait relative to a unit*. A *scale* is a linear continuum partitioned into equal units which provides the measurements, and scaling is the process of locating an entity on such a scale. Note that the term *scale* is sometimes used in social measurement in a way not consistent with *assessment*, *measurement* and *scale* as defined in this book. For example, according to the above definitions, the Likert scale and Wechsler Adult Intelligence Scale (WAIS) are not scales but assessments.

A Model of Measurement

This book is concerned with RMT and the Rasch model, a mathematical model of measurement. The theory is concerned with the approach to constructing measurements in the social sciences and goes beyond the application of the Rasch model. The Rasch model represents the structure that responses from assessments should have before they can provide measurement and how they can be transformed to provide measurements. In anticipation of studying this structure, the requirement is that within a frame of reference of assessment, which includes classes of persons and classes of items that are brought together, the *comparison* between the properties of any two persons should be equivalent no matter which subset of items is used for the comparison, and the comparison between the properties of any two items should be equivalent no matter which subset of persons is used for the comparison.

We see this structure as necessary to provide measurements. The model provides a criterion for measurement and when the responses fit the model, the requirements of measurement have in principle been met. However, we shall see that fit is not enough, and that because it is possible to obtain fit when in fact no reliable and valid measurement has taken place we must consider fit to the model carefully. The fit arises from the quality of the assessments, and we will see there is an intimate connection between the construction of the assessments and the fit of responses to the Rasch model.

The use of the model as a necessary criterion for measurement is different from the use of many statistical models which only *describe* responses (Andrich, 2004). It is part of the Rasch measurement theory. We consider this difference in the use of models more closely in subsequent chapters.

Exercises

Categorize each of the following as either nominal, ordinal, interval or ratio use of numbers according to Stevens (1951):

a. The numbers on a set of training weights in a gymnasium.
b. The numbers on a team of soccer (English football) players' shirts.

c. Scores on a biology test.
d. First, second and third place in an Olympic swimming race.
e. The numbers on a thermometer.

References

Alder, K. (2002). *The measure of all things: The seven- year odyssey and hidden error that transformed the world*. New York: Free Press.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Stevens, S. S. (1951). *Handbook of experimental psychology*. New York: Wiley.

Further Reading

Glass, G. V., & Stanley, J. C. (1970). Chapter 2: Measurement, scales, and statistics. *Statistical methods in education and psychology* (pp. 7–25). Upper Saddle River, NJ: Prentice Hall.